UNIVERSITÀ DEGLI STUDI DI UDINE

Dottorato di Ricerca in Scienze e Biotecnologie Agrarie
Ciclo XXV
Coordinatore: prof. Mauro Spanghero

TESI DI DOTTORATO DI RICERCA

Detection of somatic variants from next-generation sequencing data in grapevine bud sports

DOTTORANDA

Mara Miculan

SUPERVISORE

Prof. Michele Morgante

ANNO ACCADEMICO 2012/2013

Contents

I Summary

The grapevine (*Vitis vinifera*) is one of the oldest and the most valuable horticultural crops. Sexual crossing has been a major driver of grapevine evolution and, more recently, it has generated thousands of varieties. Somatic variation plays a crucial role in intravarietal grapevine diversity, generating novel interesting phenotypes. Somatic mutations that accidentally happened in buds of vegetatively propagated varieties were frequently noticed and the resulting bud sports were selected for their distinguished phenotype. In this work, we aimed to explore clonal variability to identify DNA mutations and transcriptional changes among genomes within a grapevine variety. *Vitis vinifera* is an ideal model because there are many clones with visible phenotypic differences and a high quality reference sequence is available (Jaillon et al 2007).

Previous studies of clonal diversity used SSR and AFLP markers that only enabled the identification of a limited number of clones. Thus, we adopted a whole genome scan approach. Illumina next generation sequencing technology was used to resequence four 'Pinot' clones ('Pinot blanc', 'Pinot gris', 'Pinot Meunier' and 'Pinot noir') and two 'Sangiovese' clones (commercially called 'R24' and 'VCR23'). Post-processed paired-end reads (2x100bp) were mapped against the PN40024 reference genome obtaining a depth of coverage >35x. Four libraries were of high quality, while the distribution of 16-kmers occurrences in the 'Pinot gris' and 'Pinot noir' Illumina reads revealed low complexity of the library and suggested to discard those clones for subsequent analyses.

SNPs were first detected in the pairwise comparison 'Pinot blanc' and 'Pinot Meunier' using the GATK – UnifiedGenotyper tool with default parameters, followed by a quality filtering and by a calibration step. In the filtering step, we removed SNPs in repetitive regions, transposable elements, and regions surrounding microsatellite motifs and INDELs, we removed bad quality SNPs based on GATK internal parameters, SNPs with <0.2 minor allele frequency and positions

with low or high coverage (<0.5-fold and >3-fold the average coverage). The calibration step was based on quality scores of a known somatic variation in 'Pinot Meunier' in the position chr1:4,897,066 (Boss and Thomas 2002). In the comparison between 'Pinot blanc' and 'Pinot Meunier' we ended up with a total of 144 putative SNPs, 79 of which were validated as true positive by Sanger resequencing (29 in 'Pinot blanc' and 50 in 'Pinot Meunier') with a FDR of 0.33 and 0.24, respectively. We performed the pairwise comparison between 'Sangiovese R24' and 'Sangiovese VCR23' with the same parameters used for 'Pinot' clones, ending up with only three putative variant positions. Of these, two SNPs were validated as true positive by Sanger resequencing. In all cases, Sanger resequencing confirmed the chimerical nature of the putative somatic mutation.

Genome scanning for copy number variations larger than 25 kbp was performed by a depth of coverage (DOC) analysis and revealed only the known somatic deletion in 'Pinot blanc' in the interval chr2:14,149,000..14,250,000 as compared to 'Pinot Meunier'. The complementary approach of paired-end mapping (PEM) revealed 11 putative deletions smaller than 25kbp in 'Pinot blanc', 19 in 'Pinot gris', 15 in 'Pinot Meunier', and 5 in 'Pinot noir' as unique to each clone and not shared with a set of 20 varieties of *Vitis vinifera* analysed with the same pipeline. In the comparison of 'Sangiovese' clones, the PEM algorithm identified seven putative deletions in 'Sangiovese VCR23', not shared with either 'Sangiovese R24' or other varieties of *Vitis vinifera*. No copy number variation larger than 25 kbp was detected by a depth of coverage (DOC) analysis between 'Sangiovese' clones.

We also compared the transcriptome of different clones in order to monitor gene expression changes that could be directly or indirectly related to somatic mutations at the DNA level. We obtained RNA-seq of leaf tissues of the same 'Pinot' and 'Sangiovese' clones analysed by DNA sequencing. Furthermore for 'Sangiovese' clones, we sequenced berry transcriptomes at two developmental stages – before ripening (2 weeks after berry set) and at the inception of ripening. More than 30,000 genes were expressed in all clones of both varieties. The vast majority of the predicted genes in the grapevine genome was transcribed at detectable levels in all

organs and stages of development investigated. Under the same experimental conditions, leaf transcriptomes were much more variable in pairwise comparisons between 'Pinot' clones than between the pair of 'Sangiovese' clones. Between the clones of 'Sangiovese', the widest differentiation in terms of global transcriptome was detected in berries collected two weeks after fruit set. Genes that showed significant differences in transcriptional levels between clones were in general not correlated with the position of the DNA mutations identified by DNA sequencing.

Through the power of the Next Generation Sequencing technology we have produced a sufficient depth and breadth of sequence coverage to comprehensively discover somatic mutations that allowed us to distinguish four 'Pinot' clones and two 'Sangiovese' clones analysed in this study. At the DNA level, somatic mutations in two 'Sangiovese' genomes appeared to be more rare than those observed among 'Pinot' clones, which corresponds to a lower level of phenotypic differentiation between the two 'Sangiovese' clones and is in accord with a presumed more recent origin compared to the 'Pinot' clones. This analysis provides the first whole-genome estimation of the rate of somatic mutation in grapevine varieties.

II Introduction

The global wine industry is dominated by a relatively small number of centuries-old varieties compared to the available natural diversity. The power of a wine brand is often linked to a grape type. In classic wine making regions, policies of product quality impose the list of varieties allowed for cultivation, which is usually restricted to varieties historically grown in the area – impeding innovation. Little differentiation naturally exists within each grape variety. This diversity is usually sought by growers for differentiating their wines on the market without changing the grape variety name on the wine label. This variation occurs in somatic mutants, most commonly referred to as clones by viticulturists. Myles et al (2011) estimated that 551 (58%) of the 950 accessions in the USDA grape germplasm collection are clones of at least another accession.

Grapevines have highly variable and heterozygous genomes. Genome heterozygosity poses challenges in the preservation of the identity of a variety over time. Superior genotypes can be perpetuated true-to-type only by vegetative propagation, which has been achieved by cutting since the antiquity or by grafting in the post-phylloxera era (Pelsy 2010). The shoot apical meristem (SAM) is organised into two histogenic cell layers. Cells in one layer proliferate independently from those in the adjacent layer. Lateral organs maintain the same cell layered structure. The outer cell layer contributes to outer and inner epidermis in the ovary and to the epidermis in the leaf lamina. The inner cell layer contributes to the formation of carpel walls, gametes, and embryo sacs in the ovary, mesophyll in the leaf, and the entirety of tissues in wood and adventitious roots.

Due to the stratified organisation of cell layers, a somatic mutation could remain restricted in the meristematic cell layer in which the mutation occurred (chimera). In a periclinal chimera the inner and the outer cell layers have distinct genotypes. Plants

that are regenerated through bud organogenesis are themselves made up of chimerical tissues and organs. Only somatic mutations that occur in the inner layer are heritable, since only this cell layer contributes to the formation of gametes. Somatic mutations that occur in the outer layer can only be fixed by vegetative propagation and deployed on a commercial scale by bud organogenesis. Less frequently, cells with mutated DNA could entirely displace wild-type cells from the meristem. The direction of this cell displacement is usually outwards, with the inner cells dividing periclinally and replacing the outermost layer. Through this mechanism, the genotype of somatic mutation originally occurring in cells of the inner layer may become homogenised in the SAM, and the plants regenerated by bud organogenesis do contain only cells containing the mutated DNA (bud sports).

II.1 Extent of the phenomenon and its impact on viticulture

Mutation rate in somatic tissues of grapevine is completely unknown on a genome-wide scale. It is assumed that the older the original seedling and the wider the acreage of planting, the higher the chance that present-day vines are the result of independent and/or sequential somatic mutations. The oldest varieties provide us with plenty of examples of chimeras and bud sports. The number of bud sports with diversified phenotypes that were selected from 'Chasselas', 'Pinot', and 'Traminer' is particularly high compared with other varieties of comparable diffusion, which lends support to the long history of cultivation of these varieties. The accumulation of somatic mutations has occurred to such an extent that some bud sports have been mistakenly considered as distinct varieties (e.g. Chasselas doré, Chasselas musqué, Chasselas sans pepins, 'Pinot blanc', 'Pinot Meunier', 'Gewürztraminer', 'Traminer rot'). In a vast survey of clonal diversity among seven famous varieties, 'Traminer' was the monozygotic group with the highest diversity in microsatellite alleles (Pelsy et al 2010). Among 59 clones of 'Cabernet Sauvignon' sampled in seven countries (France, Chile, Spain, Australia, Hungary, USA and Italy) by Moncada et al (2006), 22 different genotypes have emerged from the analysis of 84 microsatellites within a

overall genetic identity of 97%, which is compatible with the hypothesis of all of them being descended from a single seedling and the recorded diversity being due to somatic mutations. Only two clones in France and Australia carried the ancestral genotypes inferred from the parental varieties of the 'Cabernet Sauvignon' seedling ('Cabernet Franc' and ' Sauvignon') and one clone from France had the highest number of variant loci (five).

The variety 'Sangiovese' has been grown in Central Italy and Corsica for several centuries. Many synonyms and biotypes do exist as well as cases of accessions incorrectly assigned to the variety 'Sangiovese' (Di Vecchi-Staraz et al 2007).

The variety 'Sangiovese' is regarded as an ancient wine grape. The variety was first referred to as 'Sangiogheto' at the end of XVI century (Soderini 1590), but its origin is presumed to be much older. 'Sangiovese' is cultivated on approximately 11% of the Italian vineyards [Fifth General Census of Agricolture 2000] and in central Italy it is used for the production of famous red wines, such as 'Chianti', 'Chianti Classico', 'Brunello di Montalcino', 'Morellino di Scansano', 'Carmignano', 'Rosso Piceno Superiore', 'Sangiovese di Romagna'. Although the origin of 'Sangiovese' has been thoroughly investigated, the parentage remains partially unclear and debated. Microsatellite DNA of 'Sangiovese' and many National varieties suggested two hypothesis about the pedigree of 'Sangiovese'. According to one hypothesis 'Ciliegiolo' and 'Calabrese di Montenuovo' are the parents of 'Sangiovese' (Vouillamoz et al 2007). 'Ciliegiolo' is a red variety used in central Italy and frequently mistaken for 'Sangiovese' due to their highly similar ampelometric features, while 'Calabrese di Montenuovo' is a minor variety locally grown in restricted areas in Southern Italy. Contrasting molecular evidence has led Di Vecchi-Staraz et al (2007), Cipriani et al (2010), and Lacombe et al (2012) to propose the alternative hypothesis that 'Ciliegiolo' is an offspring of 'Sangiovese' and one parent of 'Sangiovese' remains unknown. A recent study of Bergamini et al (2012) claimed that an ancient variety from Southern Italy named 'Negrodolce' could be one of the parents of 'Sangiovese'.

There are over 90 approved clones of 'Sangiovese' in Italy, for most of which DNA analyses have confirmed their monozygotic origin (Filippetti et al 2005), including

those used in the present thesis (R24 and VCR23). A phyllometric analysis of 12 certified clones allowed Silvestroni and Intrieri (1995) to differentiate some biotypes, but molecular analysis with AFLP markers failed to provide strong molecular evidence for distinguishing biotypes (Filippetti et al 2005). The wide diffusion and the high intravarietal variation of 'Sangiovese' has attracted a lot of interest in clonal selection of particular variants with distinctive ampelographic and enological characteristics. Once interesting biotypes are selected and homologated, constitutors and nurseries have economic interests in their protection of plant variety rights and in the certification of the identity of clones through DNA fingerprinting.

II.2 Bud organogenesis, somatic mutations, and DNA typing of somatic chimeras

Historically, once a grapevine seedling with superior quality become noticed, its vegetative propagation secured the maintenance of genetic consistency and varietal identity over time. A novel grape variety may only originate from sexual reproduction, while somatic mutations that occur in meristematic cells give rise to phenotypic variants within a variety, thereby fixed and perpetuated by vegetative multiplication. The common sense of viticulturists in distinguishing varieties from clones was challenged in the past by borderline cases. Somatic variants may phenotypically diverge from the mother plant such extremely that the vegetative material propagated thereafter was mistakenly considered a distinct variety, as opposed to the vegetative material that conserved the characters of the original stock. This occurred to many bud sports of the glabrous-leafed, black waxed- and compact-berried 'Pinot noir' that were elevated to the rank of variety, such as the yellowish-berried 'Pinot blanc', the red-grayish berried 'Pinot gris', the hairy-leafed 'Pinot Meunier', the unwaxed-berried 'Pinot moure', the loose-berried Mariafeld-types of 'Pinot noir'.

In many cases mutations produce less pronounced phenotypic effects and the resulting variation still stays in the range of the original variety, but the phenotype is sufficiently different for the new clonal material to be uniquely distinguished and traded with a distinct clone name. Most mutations do not lead to phenotypic variants, and this represents the hidden genetic variation within varieties.

The estimation of the lowest genetic distance between varieties of sexual origin and the highest distance between vegetatively propagated-material within a variety would allow to define a measurable boundary between the width of clonal variation and the concepts of distinctness, uniformity and stability (DUS) to establish varietal identity. In a very large grape collection (Lancou et al 2011), every single accession corresponding to a different variety diverged from the others by at least four alleles over 20 microsatellite loci, even in the event of varieties deriving from self-pollination of another variety or being full-siblings. Clones could be differentiated based on microsatellite profiles only in 5 % of the cases and, if so, they differed by one to three alleles over 20 microsatellite loci. 'Orbois blanc' and 'Orbois rose' were the only pair of somatic mutants that differed by four alleles – to the same extent of differentiation as did pairs of varieties that are selfed-parent/offspring or full-siblings. The observed molecular differences between clones included changes in the heterozygous versus homozygous status – probably caused by mutations in the primer annealing sites and allele drop or deletion of the entire locus – as well as shifts in size of one allele, or chimerical triallelic profiles generated by shift in size of one allele in a confined portion of the meristem.

II.3 Definition of variety and clone in viticulture

A variety derives from sexual reproduction and all plant material assigned to a varietal name is derived from a initial single monozygotic seedling. All clones within the variety have the same monozygotic origin. Sometimes, a single mutation has a dramatic phenotypic impact (**Figure 1**). Although the derived material is by definition a clone, the new material has been sometimes considered as a essentially derived

variety (EDV), as it occurred to the 'Pinot' variants. All registered clones of 'Pinot blanc' investigated by Vezzulli et al (2012) are bud sports and originated from the same deletion in the DNA of 'Pinot noir'. Clones of 'Pinot gris' are chimeras that originated from an independent deletion in the inner cell layer of the meristem in 'Pinot noir'. White-skinned clones of 'Pinot' also arose from two successive layer-specific mutations in the DNA of 'Pinot noir', through the intermediate state of chimerical 'Pinot gris' (Furiya et al 2009). By contrast, the white-skinned 'Cabernet Sauvignon' also known as 'Shalistin' has derived from one single layer-specific mutation, which gave rise to the chimerical 'Malian' – the counterpart of 'Pinot gris' – followed by a event of cell rearrangement by which the mutated cells of the inner layer displaced the wild-type cell from the outer layer (Walker et al 2007). It remains unknown whether 'Pinot blanc' has derived from two independent events of deletion that successively occurred to the red allele in the inner and the outer layers of the meristem in 'Pinot noir' or from a single mutational event in the inner layer followed by cell displacement into the outer layer.

Figure 1- Berry color mutants of 'Pinot' clones . Left to right: 'Pinot noir' (black), 'Pinot gris' (grey and 'Pinot blanc' (white) (This P et al, 2006)

II.4 DNA changes in somatic mutants

DNA polymerase slippage, single nucleotide substitutions, mobilisation of transposable elements, and large chromosomal deletions are the documented causes of DNA changes in somatic mutants of grapevine.

II.4.1 Microsatellite DNA variation

DNA polymerase slippage in tandemly repeated short motifs is a prominent source of DNA variation between cell populations in somatic mosaics and among clones, though phenotypically silent in all cases documented so far. This became evident since the early 2000s (Riaz et al 2002; Franks et al 2002). The interrogation of less than a hundred randomly selected loci was enough to disclose DNA variation among some but not all clones of 'Cabernet Sauvignon' and 'Pinot noir' (Hocquigny et al 2004, Moncada et al 2006). This variation occurred in intergenic regions and was not associated with noticeable phenotypic diversity.

II.4.2 Single nucleotide substitutions

A single nucleotide mutation that occurred in the outer cell layer of the meristem and caused an amino acid substitution in the DELLA domain of the giberellic acid-signalling protein GAI1 is the DNA variant that differentiates the hairy-leafed 'Pinot Meunier' (**Figure 2**) from the original genome of 'Pinot noir'. When mutated cells were isolated from the chimeric apex of 'Pinot Meunier' and used for somatic embryogenesis, the regenerated non-chimeric *GAI1* mutants are dwarf and hyper-fruity, as a result of impaired giberellic acid signalling and abnormal floral induction opposite to each leaf (Boss and Thomas 2002, Franks et al 2002).

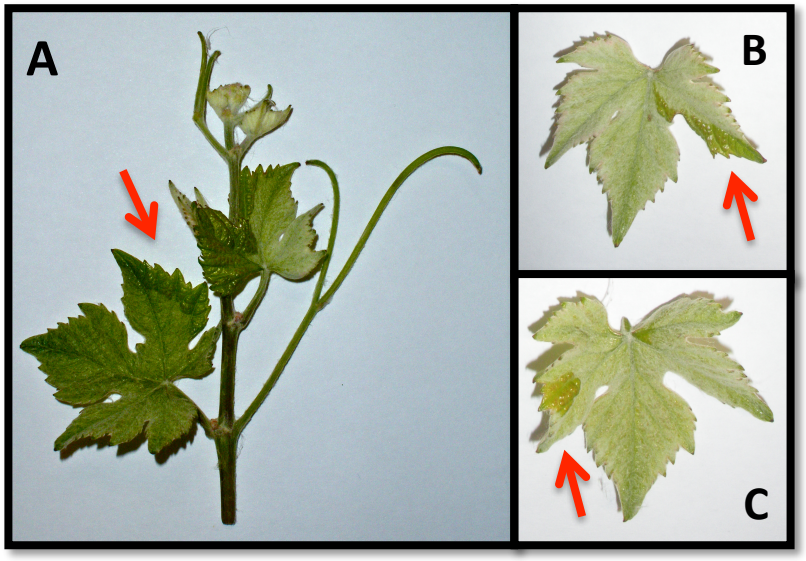


Figure 2 – Examples of periclinal chimera 'Pinot Meunier' shoot (A) and leaves (B-C). The variant nucleotide in position chr1:4,897,066 of L1 cell layer is responsible of the tomentose phenotype of 'Pinot Meunier' leaves. Red arrows indicate hairless sectors where the L2 cell layer has displaced the L1 cell layer.

II.4.3 Transposable elements

Mobilisation of transposable elements from and into the coding or the promoter region of functional genes is a frequent cause of phenotypic changes associated with somatic mutations.

Inactivation of the MybA1 transcription factor in the white-colour haplotype of the berry skin colour locus in grapevine was caused by the insertion of a TE that disrupted the MybA1 promoter region. Excision of the LTR-retrotransposon *Gret1* from the promoter region of MybA1 partially restores the expression of the downstream gene, triggering anthocyanin biosynthesis. Intra-LTR recombination is the most frequent mechanism of excision, which leaves a solo-LTR as a footprint of the former presence of the TE. This somatic mutation occurred independently in 'Rubi', 'Benitaka', 'Ruby Okuyama', which all derived from the white-skinned 'Italia', and in 'Flame Muscat' – a bud sport of 'Muscat of Alexandria' –,'red Chardonnay' and 'pink Sultana'. Something similar occurred to the 'Traminer' plant that originated the pale-colored 'Gewurztraminer', in which a small 44-bp insertion is the footprint of *Gret1* insertion and excision.

Insertions of TE in promoter regions not only cause can cause inactivation of the downstream gene but they can otherwise promote exaggerated or ectopic expression.

Transposon-mediated *cis*-activation of a gene homologous to the *Arabidopsis* TERMINAL FLOWER 1 (TFL1) – involved in inflorescence development – is responsible for the reproductive meristem (RRM) phenotype of the RRM somatic mutant of Carignan. In that case, the insertion of a class II DNA transposon in the promoter of *TFL1* enhances *TFL1* expression and causes exaggerated proliferation and branching of the inflorescence, resulting in huge clusters, along with minor alterations in flower morphology and delayed anthesis (Fernandez et al 2010).

In a similar way a miniature inverted-repeat transposable element insertion in the promoter region of the *PISTILLATA*-like (*VvPI*) gene causes the ectopic expression in the fruit of genes specific for petal and stamen development. This alters the cell differentiation patterns in the ovary, impairing normal development of berry flesh in the FLESHLESS BERRY (FLB) somatic variant of the variety 'Ugni Blanc' (Fernandez et al 2012).

II.4.4 Structural variation

A structural deletion of a block of DNA amounting to 100-170 kb is responsible for the complete elimination of the MybA gene cluster from the once red allele of 'Pinot noir' that is today found in its derived white-fruiting bud sport 'Pinot blanc' (Yakushiji et al 2006, Vezzulli et al 2012). A even larger deletion of approximately 4 million nucleotides has eliminated one quarter of chromosome 2 and approximately 200 genes in the inner cell layer of 'Pinot gris' (Vezzulli et al 2012).

II.4.5 Other changes

In addition to these well-documented cases, observations in somatic variation of human cells provide arguments that clonal variants in plants may also arise from copy number variation, epigenetic modifications, and misregulation of microRNA and

small RNA pathways, though experimental evidence for these mechanisms has yet to come in grapevine.

II.5 Strategies to identify DNA changes in somatic mutants

II.5.1 Mutations in candidate genes associated with a mutated phenotype

Clonal variation in specific traits with known genetic control can be investigated by a DNA analysis targeted to the region of interest. This is the case of berry colour mutants that are easily differentiated by inspecting DNA sequence at the MybA gene cluster, a pair of closely linked transcription factors involved in the control of color (Walker et al 2007).

Fruit-colored bud sports that reverted from white fruiting varieties and accumulate anthocyanins in berry skin are the most common case of somatic mutation. Mutations are caused by independent events of partial excision of the Class I LTR-retrotransposon *Gret1* from the promoter region of the MybA1 transcription factor by intra-LTR recombination, which partially restores expression of the transcription factor, thereby leading to the synthesis of the key enzyme for anthocyanin biosynthesis (reviewed in Pelsy 2010).

II.5.2 Untargeted approaches based on DNA marker screening on a whole genome scale

Traditionally, untargeted approaches to the scanning of DNA variation from grapevine clones has been done through the generation of AFLP, other PCR-based and retrotransposon-based markers. AFLP are claimed to generate polymorphic banding patterns that distinguish most of the clones within a variety and even reveal association with the geographical origin of the vegetatively propagated material (Meneghetti et al 2011). Other authors are most sceptical in the use of AFLP and other PCR-based markers due to low repeatability of the banding patterns until they

become converted and validated into SCAR markers. In our laboratory at Institute of Applied Genomics (IGA), we performed a test in which differential bands were generated from the DNA from different clones, but these differences were inconsistent among independent sampling and DNA extractions (unpublished data).

II.5.3 Novel approaches enabled by Next-Generation Sequencing technologies

The field of DNA sequencing has undergone rapid advances (Clyde 2007). The overwhelming majority of DNA sequencing to date has relied on evolving versions of the Sanger biochemistry and technologies of electrophoresis (Deschamps and Campbell, 2010). The application of automated Sanger sequencing for genome analysis is considered to be the first-generation technology and despite many technical improvements, there were some insurmountable limitations that revealed a need for new approaches that are referred to as next-generation sequencing (NGS). NGS relies on a combination of breakthrough methods of template preparation, chemistry of sequencing, imaging, and bioinformatics methods of genome alignment and assembly (Metzker 2010).

The advent of NGS platforms have drastically increased the speed at which DNA sequences can be acquired, the ability to produce an enormous volume of data and reduced the costs by several orders of magnitude. In summary, the arrival of NGS technologies has changed the way to think about scientific approaches in basic, and applied research. The ability to sequence the whole genome of many related organisms has allowed large-scale comparative and evolutionary studies to be performed that were almost impossible with previous technologies.

The reduction in read length and quality has required the development of bioinformatics tools to assist in the mapping of shorter reads to the reference genome and in the *de novo* assembly of entire genomes. NGS platforms offer much higher throughput with greatly reduced costs but with lower accuracy: the error rate is 10-fold greater than the one obtained by Sanger Sequencing (Shendure and Ji 2008). In order to compensate for the lower quality, the high throughput provides a

redundancy of reads at a given nucleotide position (coverage), which can be employed to discern sequencing errors from true genetic variation (Hillier et al 2008).

NGS technologies have a wide range of possible applications, and more are being developed. Current applications include:

i) full genome re-sequencing or variant discovery by resequencing of targeted regions of interest among individuals;

ii) de novo assemblies of bacterial and lower eukaryotic genomes;

iii) mapping of structural rearrangements, which may include copy number variation, balanced traslocations breakpoints and chromosomal inversions within a population;

iv) RNAseq to measure gene expression and to catalogue the transcriptomes of cells, tissues and organisms, to elucidate the role of non- coding RNAs in health and disease (libraries derived from mRNA, totalRNA or smallRNA are deeply sequenced) Wold et al 2008; Wang et al 2009; Ponting et al 2009)

v) large-scale analysis of DNA methylation (epigenetic marks); v) ChIP-seq, or genome-wide mapping of DNA-protein interactions and chromatin structure, by deep sequencing of DNA fragments pulled down by chromatin immunoprecipitation (Wold et al 2008).

vi) species classification and/or gene discovery by metagenomics studies (Petrosino et al 2009).

In addition to the applications described above, NGS technologies are being used to characterize the evolutionary relationships of ancient genomes.

II.5.4 Illumina NGS technology

One of the available technologies for NGS is the so-called Solexa-Illumina or for the sake of simplicity Illumina. In common with other technologies, the Illumina protocol for library preparation can be substantially summarized in three steps (**Figure 3**): i) random fragmentation of nuclear acid material, either via nebulization or sonication;

ii) ligation of universal adapters at both ends of the fragmented DNA/cDNA; iii) immobilization and amplification of the adapter-flanked fragments to generate clustered amplicons that will be the templates for the sequencing reaction.

Single molecules are covalently attached to a planar surface and amplified in situ. Sequencing by synthesis is carried out by adding a mixture of four fluorescently labelled reversible chain terminators and DNA polymerase to the template and by adding a single reversible terminator to each template. The fluorophore and the reversible block are removed after the detection of the fluorescent signal for each template. The terminator–enzyme mix is then added to start the next cycle, and the process is reiterated until the end of the run. The sequence accuracy is guaranteed by the presence of all the four in the reaction, minimizing the risk of mis-incorporation. Accuracy is also independent of sequence context, and a discrete signal is generated for every base.

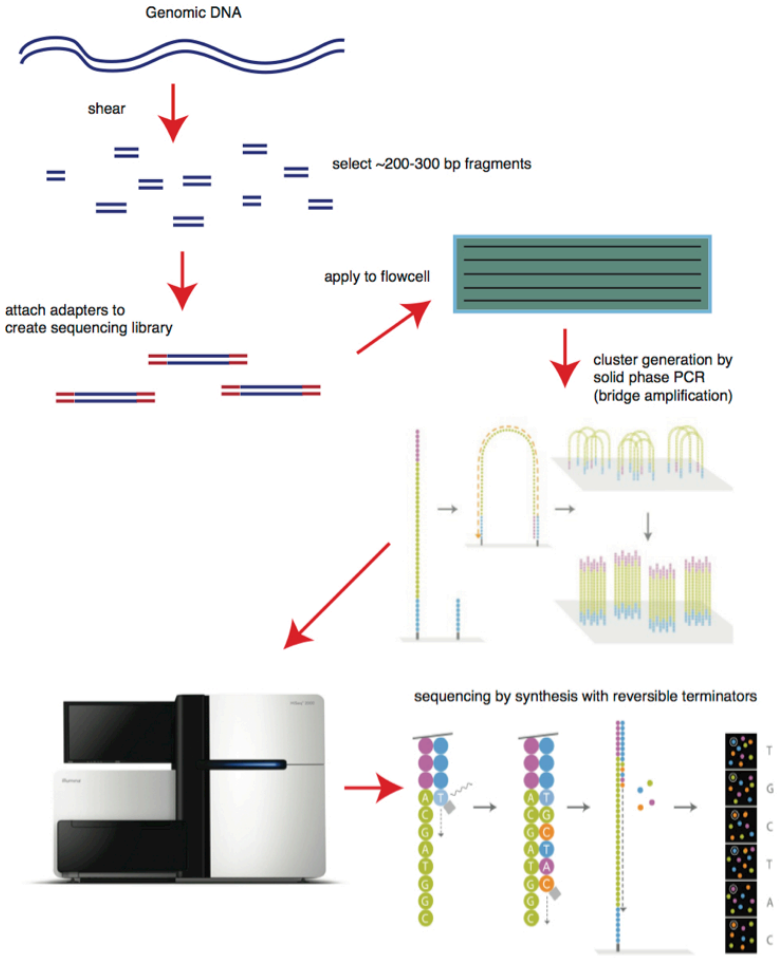


Figure 3 – Illumina sequencing technology. Genomic DNA is fragmented and gel selected at 500-600bp. Adapter-modified, single-stranded DNA is added to the flow cell and immobilized by hybridization. The technology relies on bridge PCR to amplify clonal sequencing features. Clonally amplified clusters are denatured and cleaved; sequencing is initiated with addition of primer, polymerase (POL) and 4 reversible dye terminators. Post-incorporation fluorescence is recorded. The fluorophore and block are removed before the next synthesis cycle. Accurate measurement of the concentration of the template library is critical to maximize the cluster density while simultaneously avoiding overcrowding (Shendure et al. 2008).

II.5.5 Untargeted approaches in the genome sequencing era

The nuclear genome of a model grapevine genotype has been entirely assembled, decoded, and released in 2007 (Jaillon et al 2007). The reference sequence offers the framework against which to compare any other variety in parallel analyses of genome-wide polymorphisms. DNA short-reads are generated from other grapevines, using one of the available instrument platforms for NGS, and aligned against the assembled reference sequence (Mardis et al 2011). Bioinformatics

algorithms and pipelines are then utilized for calling single nucleotide polymorphisms (SNPs), insertions/deletions (indels), and structural variants of individual genomes. At the time the first NGS analysis was performed in grapevine (Myles et al 2010), DNA reads were not longer than 36 bp and reduced representation libraries were sequenced, compounding the complexity of sequence alignment. Despite these limitations, thousands of SNPs were discovered among eleven varieties. A similar approach was applied to the analysis of clonal diversity by Carrier et al (2012) using short reads produced by the Roche 454 GS FLX technology at a rather low coverage (less than 1X).

Large structural variation and copy number variants are usually scanned through an entire genome by mapping NGS reads and detecting aberrations in depth-of-coverage.

DNA typing of plant somatic mutants is as challenging as sequencing cancer genomes in human tumor tissues. Any tissue section normally selected for genomic DNA isolation (leaf, berry, flower) will include normal cells and mutated cells in unpredictable proportions, which causes overlapping between normal DNA signatures and altered signatures provided by the population of mutated cells.

II.5.6 Transposon insertion-site profiling using NGS

Most of the known phenotypic variation between grapevine clones is accounted for by the activity of transposable elements (TE). This evidence emerged from the investigation of specific gene regions already known to be responsible for trait variation (i.e. *MybA*) or transcriptionally altered in mutants (i.e *TFL1*). These observations have generated the expectation that transposition of Class I and II mobile elements be the most frequent cause of somatic mutations.

Systematic monitoring of transposon activity is today possible and relatively easy with the use of NGS. The earliest application of NGS to address the issue of clonal variation made use of gapped alignment of Roche 454 GS FLX single reads, averaging 355 bp in length, from clones of ‘Pinot noir’ (Carrier et al 2012). Estimates indicate a

frequency of 35 TE polymorphic sites per million nucleotides among clones compared to as few as 1.6 SNPs and 5.1 small indels per million nucleotides. Most of this variation is supposed to be phenotypically silent, but more than half of those events were localised in genic regions and may point to some biological role worth to test. Shorter but paired sequences are now available from both ends of sheared and size-selected fragments using Illumina and Applied Biosystems NGS platforms. This sort of reads can be handled by different bioinformatics tools to scan the genome for deletions and novel (non-reference) TE insertions: (i) gapped alignment of reads or so-called split-read method, (ii) deviation from mean library insert-size in paired-end mapping, and (iii) anchorage to a single genome region of unpaired reads orphaned by nonalignment of their mates, which are then assembled into sequence contigs and compared to TE databases. A genome-wide transposon insertion-site profiling may become the method of choice for the systematic scanning of somatic variation (Baillie et al 2011)

II.5.7 Somatic mutations in Cancer Genomes

The somatic mutations concept is not only relevant in plant genomes. In fact, the genomes of all cancer cells, and indeed of most normal cells, have acquired a set of somatic mutations, independent from germline mutations. Some of these somatic variations confer selective clonal growth advantage and are causally implicated in oncogenesis, being positively selected during the evolution of the cancer. For this reason such mutations are know as 'Driver mutations' (Stratton, et al, 2009). Somatic variations involved in cancer causation include point mutations, genomic rearrangements and changes in copy number. Over the last thirty years several strategies have been used to detect the various classes of somatic mutations in cancer genomes such as G-banded cytogenetics, spectral karyotyping, FISH, and copy number arrays (Campbell et al, 2008; Pleasance et al, 2009). Nevertheless, these strategies are unable to detect anything less than gross genomic rearrangements, provide limited resolution of breakpoint mutations and do not report on balanced

rearrangements or fusion events. Both end-sequencing of BAC libraries built from cancer genomes and hybridization of flow-sorted chromosomes to arrays are not applicable to large numbers of cancer genomes. Next Generation systematic sequencing offers therefore the potential to carry out a genome-wide screening of all somatic mutations of all classes in individual cancer genomes, leading up to the characterization of a complete catalogues. Lung cancer, malignant melanoma and a lymphoblastoid cell line were the first cancer lineages to be extensively investigated with Next Generation Sequencing (Campbell et al, 2008; Pleasance et al, 2009).

The International Cancer Genome Consortium (ICGC, http://icgc.org) was created to comprehensively characterize somatically acquired genetic events in at least fifty classes of cancer.

III Objectives of this thesis

We primarily aim at setting up a bioinformatics pipeline that makes use of Next Generation Sequencing data to detect somatic variants in grapevine bud sports. To this end, we planned to re-sequence with Illumina technology four clones of 'Pinot', for some of which known DNA mutations are documented in literature reports, and two outstanding clones of 'Sangiovese' for which the wine industry and grapevine nurseries would wish to certify the genetic identity of the commercialised plant material. We intended to use known mutations to calibrate the procedure and then apply it to variant discovery in 'Sangiovese'.

NGS also offers new opportunities for transcriptome analysis. Deep sequencing of mRNA (RNA-Seq) allows to discover novel transcripts without any a priori knowledge of the genes and to measure transcript levels of all genes in a single assay. Most importantly for the scope of this thesis, RNA-seq allows to monitor gene-specific expression changes that could be directly or indirectly related to somatic mutations at the DNA level. In this work, we planned to perform RNA-seq of leaf tissues in the same clones of 'Pinot' and 'Sangiovese' used for DNA sequencing. This analysis has the dual scope of assessing the changes in global gene expression among somatic mutants and of detecting specific gene expression changes that could be directly relates to DNA changes. For the 'Pinot' clones carrying known mutations, we also aim at evaluating the global impact imparted by these mutations in the leaf transcriptome, since most of the past studies focused on berry transcriptome alone.

For the clones of 'Sangiovese' we also planned to sequence berry transcriptomes at two developmental stages – before ripening (2 weeks after berry set) and at the inception of ripening (80% of coloured berries over the clusters) to monitor transcriptional changes between clones.

IV Results

IV.1 Resequencing and Reference mapping

IV.1.1 Filtering and alignment of Illumina reads

The raw FASTQ data for two clones of 'Sangiovese' and four clones of 'Pinot' were processed for adapter removal, quality trimming and filtering for contaminants and duplicates. Post-processed paired-end reads were aligned to the reference genome of PN40024 using BWA (Li and Durbin, 2009) with default parameters. Metrics of this process are given in Table 1.

IV.1.2 Distribution of genome coverage

The uniformity of genome coverage was estimated by counting the occurrence of 16-kmers in the Illumina reads. The graphs are reported in **Figure 4**. The libraries obtained from the 'Sangiovese' clones, 'Pinot Meunier' and 'Pinot blanc' displayed the expected distribution with three discernable peaks in genome coverage (that are normally observed also in other grapevine genotypes), presumably corresponding to the diploid genomic regions, the haploid ones (hemizygous/heterozygous DNA), and the duplicated ones, respectively. Despite the high number of reads obtained from 'Pinot noir' and 'Pinot gris', the distribution of kmers seems to indicate a low complexity in the library, which suggested us not to use these sequences in subsequent analyses.

Table 1- Metrics of the Illumina reads from raw reads to final coverage of unique aligned reads

	raw reads	bp before trimming and filtering	raw coverage	trimmed and filtered reads	bp after trimming and filtering	mapped reads	mapped reads - unique	removed duplicates	bp covered in the reference - unique	% of the reference genome covered - unique	coverage of unique reads
Pinot blanc	389,834,062	38,983,406,153	80,18	308,249,931	30,824,993,142	212,159,929	154,019,370	11,94%	437,562,933	90%	35
Pinot gris	425,229,322	42,522,932,180	87,46	398,050,818	39,805,081,838	258,954,966	203,765,098	22,83%	429,717,325	88%	47
Pinot Meunier	318,654,582	31,865,458,210	65,54	265,318,592	26,531,859,239	203,011,008	163,097,990	13,05%	439,343,510	90%	37
Pinot noir	313,500,877	31,350,087,662	64,48	226,228,223	22,622,822,254	103,169,532	90,477,516	29,70%	357,528,738	74%	25
Sangiovese R24	512,453,356	48,512,899,301	99,78	387,856,854	36,547,551,017	297,970,360	240,419,477	9,25%	434,506,311	89,40%	50
Sangiovese VCR23	474,627,103	46,402,797,247	95,44	393,950,543	38,492,345,537	297,490,884	244,411,323	10,41%	434,542,361	89,40%	53

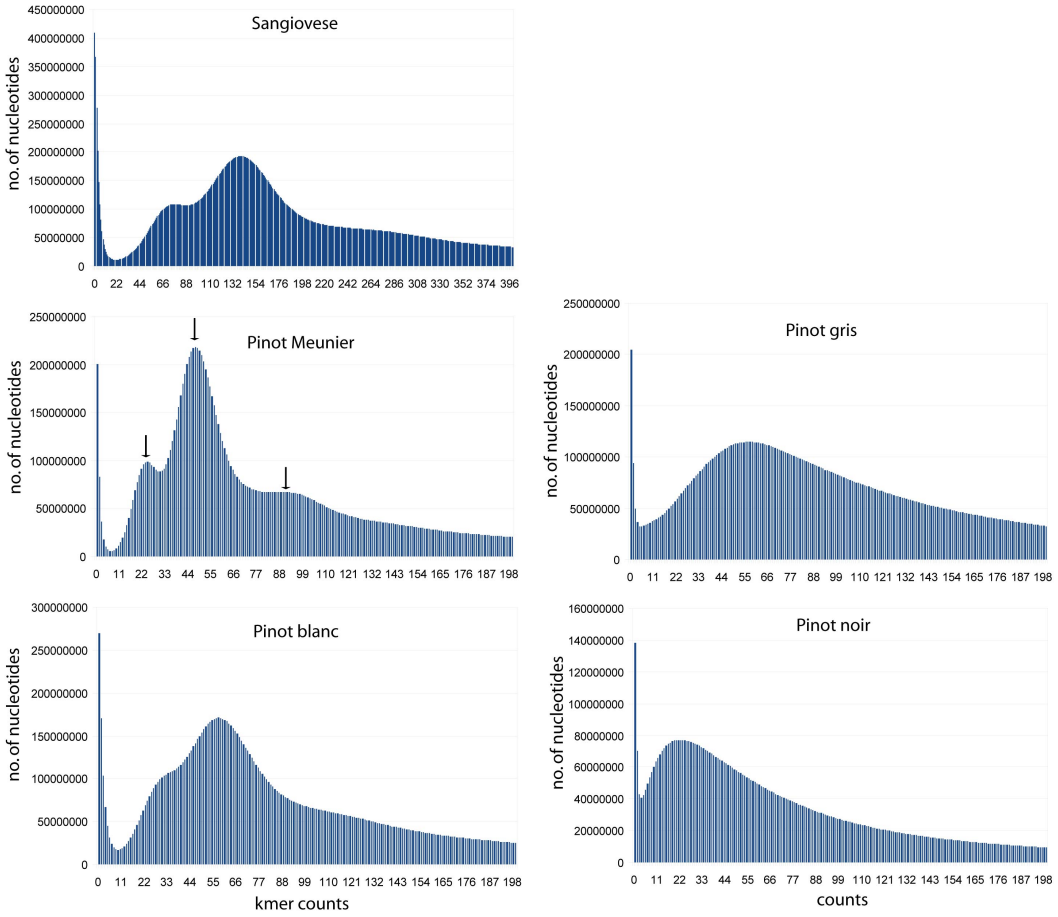


Figure 4 – Distribution of 16-mers in the Illumina reads obtained from 'Sangiovese' (reads from both clones merged for generating the graph) and from four clones of 'Pinot'. The mode is indicative of the coverage of genomic regions present in dual copy, the minor peak at approx. 0.5-fold coverage of the mode is indicative of the coverage of hemizygous regions, the peak at approx. 2-fold coverage of the mode is indicative of the duplicated fraction of the genome. The three peaks are indicated by arrows in the graph of 'Pinot Meunier'.

IV.2 SNP detection

In order to calibrate the parameters for SNP detection, we used the comparison between 'Pinot blanc' and 'Pinot Meunier', because they differ by a known nucleotide substitution and a long segment of hemizygous DNA. The documented nucleotide substitution occurred in 'Pinot Meunier' in the *VvGAI1* gene at position chr1:4,897,066 of the reference genome (Boss and Thomas, 2002). 'Pinot blanc' is homozygous 'A', while 'Pinot Meunier' is heterozygous 'AT' in a chimerical state

because the A-to-T substitution has occurred only in one meristematic cell layer and leaf tissues are composed of a mixture of 'AA' and 'AT' genotypes. This case-study represents the most disadvantageous situation for SNP detection because the variant allele is expected to be present in the variant clone at a frequency substantially lower than 0.50, in the range of frequency (0.10-0.30) of false positive SNPs (paralogous positions and sampling error in positions with low coverage). This is due to under-representation of the mutated tissues in the plant material used for DNA extraction.

The known deletion has been estimated to span ~100-179 kb on chr2:14,149,000..14,250,000. The deletion occurred in a region present in a heterozygous state in the wild-type, and the hemizygous DNA of the haplotype still present in 'Pinot blanc' is identical to the reference genome. This case-study represents the condition of loss-of-heterozygosity and it is detectable thanks to the fact that the coverage in the hemizygous region is approximately half of the average genome coverage in the variant clone, and the allelic variant in the wild-type clone is expected at a frequency of 0.50, barring sampling errors.

IV.2.1 Variant detection in 'Pinot blanc and 'Pinot Meunier'

A total of 4,618,105variable positions were detected between 'Pinot blanc' and the reference genome sequence. A total of 4,697,469 variable positions were detected between 'Pinot Meunier' and the reference genome sequence. These positions were called using the default parameters of UnifiedGenotyper of the GATK package, version 2.1-13 (McKenna A et al. 2010).

IV.2.1.1 Filtering

The set of raw SNPs was filtered against:

- variable positions shared by both clones
- variable positions in repeated regions, transposable elements and small indel/SSR intervals
- minimum coverage <0.5-fold the average coverage

- maximum coverage >3-fold the average coverage
- GATK Phred-scaled quality score (QUAL) < 100
- GATK Strand Bias (SB) > 0
- Phred-scaled likelihoods (GATK PL) for each of the 'homozygous reference', 'heterozygous', 'homozygous alternate' (respectively 0/0, 0/1, 1/1) possible genotypes: 'true genotype' = 0, others < 30, Σ other <300

resulting into a list of 20,449 filtered SNPs uniquely present in either clone.

A further filtering step for the distance from the end of the read for reads with the alternate allele (GATK ReadPosRankSumTest < -2) reduced the list of filtered SNPs to 8,958. At this stage of filtering, we checked that the known SNP at chr1:4,897,066 in 'Pinot Meunier' was retained, and we plotted the scores for parameters of SNP quality, separately for the variable positions called in the hemizygous regions around MybA and for the variable positions in the remainder of the genome.

IV.2.1.2 Calibration

The quality scores of the variant positions (**Figure 5** A) show an excess of low quality positions at the WG level when compared to the chr2:14,1..14,3 interval, but the quality score of the known variant at chr1:4,897,066 in 'Pinot Meunier' (298.99) suggested that a more stringent threshold would pose the risk of filtering out chimerical heterozygous SNPs that have a low quality score because of low frequency of the variant. The read position bias (**Figure 5**B) for high quality SNPs in the chr2:14,1..14,3 interval indicate that many false positives on a WG scale may be associated with ReadPosRankSum >2.5. The known variant call at chr1:4,897,066 in 'Pinot Meunier' has a ReadPosRankSum of 0.86 which matches quite closely the average value scored by the high quality set of SNPs in the chr2:14,1..14,3 interval. The distribution of the likelihoods for the alternative genotype calls indicated that a likelihood lower than 50 is unlikely to occur (**Figure 5**-C) and there is a excess of SNP with likelihood comprised between 30 and 50 on a WG scale. The plots of **Figure 5**-D and **Figure 5**-E confirmed that with these filtering parameters the SNPs called in

the chr2:14,1..14,3 interval conform to expectations. The SNPs called in 'Pinot blanc' display loss-of-heterozygosity and have a distribution of coverage that is approximately 0.5-fold lower than that found in 'Pinot Meunier'. The SNPs called in 'Pinot Meunier' have a distribution of variant frequency with the mode pointing to 0.5, while the chimerical heterozygous SNP at chr1:4,897,066 in 'Pinot Meunier' has a variant frequency of 0.23.

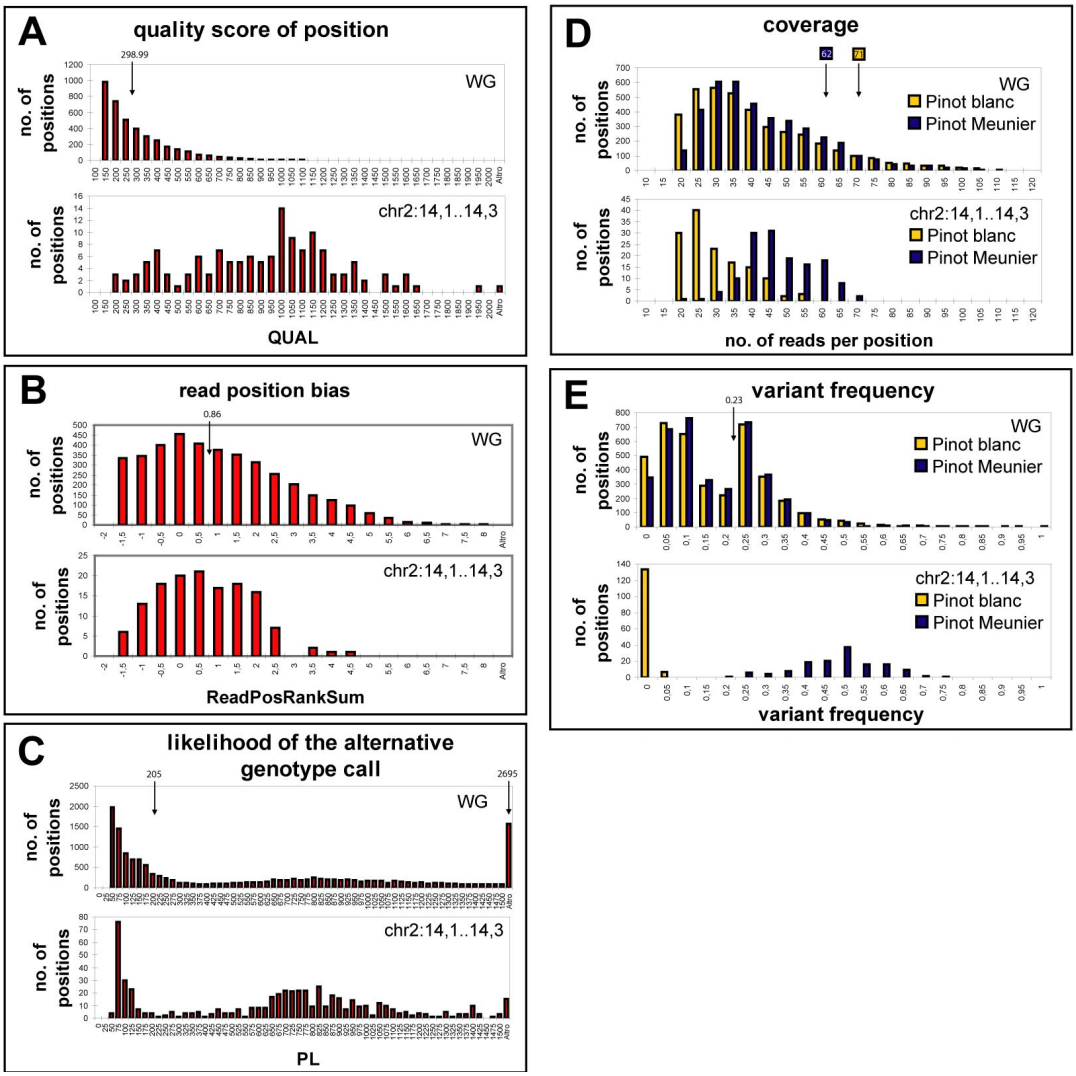


Figure 5 - Plots of scores for five parameters affecting SNP quality in the hemizygous region of chr2:14,1..14,3 Mbp versus the remainder of the genome (whole genome, WG). The scores for the known SNP at chr4:14,897,066 in 'Pinot Meunier' are indicated with an arrow for the parameter (A) quality score of the chromosomal position call in both individuals, (B) bias in the position of the variant within the reads that carry the variant, (C) likelihood of call for the two possible alternative haplotypes. Genome coverage (D) and frequency of the variant (E) are given separately for each individual. Note that SNP variant frequencies in 'Pinot Meunier' chr2:14,1..14,3 refer to

heterozygous position that became mutated in 'Pinot blanc', while SNP frequency at chr1:14,897,066 refers to a chimerical heterozygous position due to a mutation that occurred in 'Pinot Meunier'.

Based on the evidences presented above, the list of 8,958 SNPs was further filtered for:

- distance from the end of the read for reads with the alternate allele (GATK ReadPosRankSumTest > 2.5)
- Phred-scaled likelihoods (GATK PL) for each of the 'homozygous reference', 'heterozygous', homozygous alternate' possible genotypes: 'true genotype' = 0, others < 50
- adjacent SNP (< 100bp) without evidence of hemizygosity (no significant difference in genome coverage between individuals)

With this filtering, we ended up with 889 variant positions between 'Pinot blanc' and 'Pinot Meunier'. Among these, 539 were homozygous or hemizygous for the reference allele in 'Pinot blanc' and apparently 538 heterozygous and 1 homozygous variant in 'Pinot Meunier'. The remaining 350 variable positions were homozygous or hemizygous for the reference allele in 'Pinot Meunier' and apparently 348 heterozygous and 2 homozygous variant in 'Pinot blanc'. The 539 positions that were putatively mutated in 'Pinot Meunier' include the known chimerical heterozygous SNP at chr1:4,897,066 in 'Pinot Meunier' (Boss and Thomas, 2002) and 104 SNPs that fall in the 14,115,143 to 14,309,249 interval identified as heterozygous deletion in 'Pinot blanc'. With calibrated parameters of filtering, the SNPs called in the chr2:14,115,143..14,309,249 interval display the metrics shown in **Figure 6**. At this point we examined the allele frequency spectrum and filtered out positions where the minor allele frequency was < 0.2 and positions that reveal incongruence between the PL values and the genotype call by GATK UnifiedGenotyper tool, version 2.1-13 (McKenna A et al. 2010). With this filtering we ended up with 109 variant positions (107 heterozygous and 2 homozygous) in 'Pinot blanc' and 375 (374 heterozygous and 1 homozygous) in 'Pinot Meunier'.

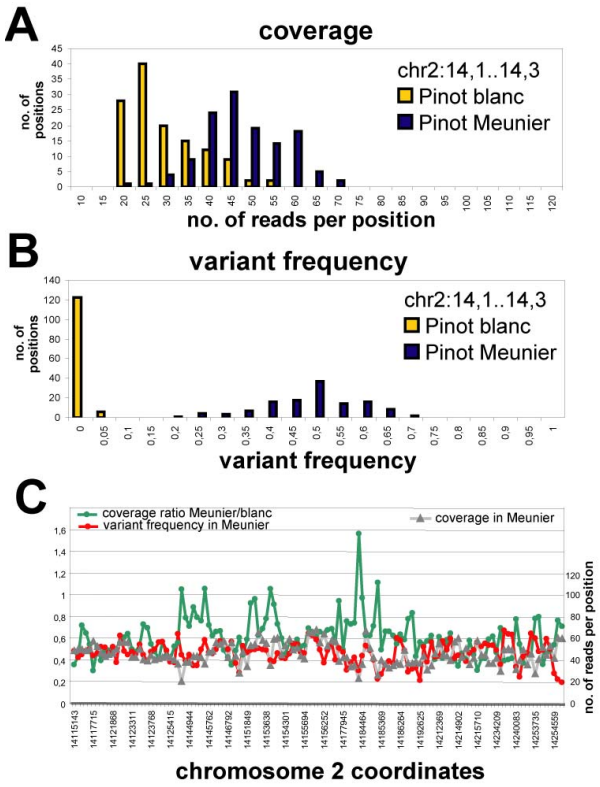


Figure 6 - Distribution of (A) coverage and (B) variant frequency of SNPs in the region chr2:14,1..14,3 Mbp. Coverage ration between 'Pinot Meunier' and 'Pinot blanc', variant frequency, and coverage in 'Pinot Meunier' (C) of the high-quality SNPs identified with calibrated parameters. Coverage in 'Pinot Meunier' is referred to the right y-axis

IV.2.2 SNP confirmation by capillary sequencing

In order to evaluate the effectiveness of our SNP detection pipeline in finding differences among somatic clones, we visually inspected the 484 variant positions and ended up with a selection of 169 unique SNPs between 'Pinot blanc' and 'Pinot Meunier' to be experimentally validated. We were able to design PCR primers across the region for 54 putatively mutated positions in 'Pinot blanc' and for 90 in 'Pinot Meunier'. We proceeded with Sanger sequencing of genomic DNA extracted from young leaves. Of these, 9 and 18 amplicons, respectively, did not produce a sufficiently clear signal, preventing the validation; of the remaining positions, 2 and 6, respectively, produced a too low signal, which prevented classification in true or false positive SNPs position. In 'Pinot blanc' and in 'Pinot Meunier' a total of 29 and

50 putative SNPs were validated as true positive single base variants indicating a FDR of respectively 0.33 and 0.24 (**Table 2**). With the exception of one SNP in 'Pinot blanc', all variant positions were heterozygous. The known somatic SNP in chr1:4,897,066 was included in the validation resulting in a true positive variation with a low signal (**Figure 7**.A). A total of 46 out of 79 SNPs are in exons or introns.

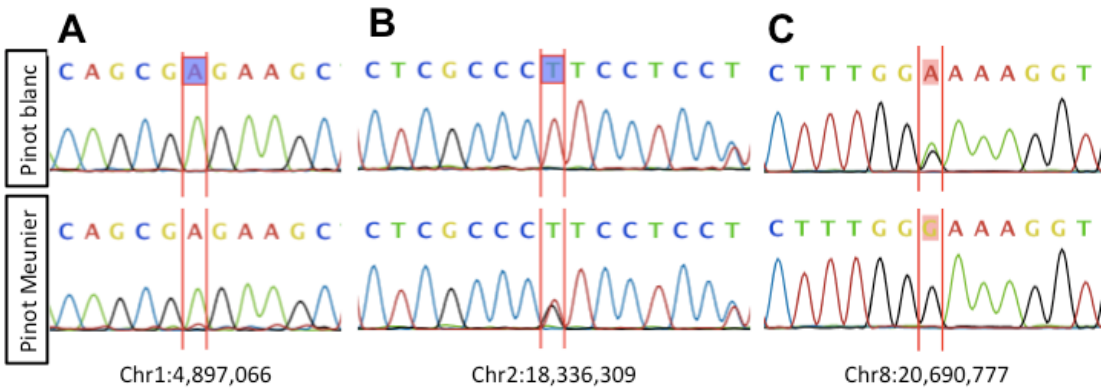


Figure 7 – Comparison of electropherograms obtained by Sanger resequencing of three regions across the putative somatic SNPs detected among the studied 'Pinot blanc' and 'Pinot Meunier'. The vertical red lines indicate the polymorphic position in L1+L2 derived tissue (leaf). In particular, (A) shows the validation of the known somatic position at chr1:4,987,066: 'Pinot blanc' electropherogram shows the homozygous state A/A, while 'Pinot Meunier'. electropherogram shows the heterozygous state with an under-represented allele A/T. (B) shows the validation of the chr2:18,336,309 for the variant position in 'Pinot Meunier' which genotype is heterozygous T/C while 'Pinot blanc' is homozygous reference T/T. In (C) there is the validation of the chr8:20,690,777 variant position for 'Pinot blanc' which genotype is heterozygous G/A while 'Pinot Meunier' is homozygous refererence A/A.

Table 2 – Genotype calls from Sanger amplicons sequenced in the experimental validation of unique variant positions selected from the somatic SNP detection pipeline in the comparison 'Pinot blanc' and 'Pinot Meunier'. The SNPs in table are the true positive variant mutations.

			PB	PM	
SNP position	**Variant clone**	**reference**	**Sanger Genotype**		**Region**
chr1:4745158	Pinot blanc	T	T/C	T/T	1.6Kb dowstream gene
chr1:7321481	Pinot blanc	C	C/G	C/C	3Kb dowstream gene
chr10:7060090	Pinot blanc	G	A/A	G/G	exon
chr11:5376149	Pinot blanc	C	T/C	C/C	intron
chr12:17004139	Pinot blanc	C	T/C	C/C	intron
chr15:10456989	Pinot blanc	C	C/T	C/C	intergenic
chr16:19659567	Pinot blanc	C	C/T	C/C	intron
chr17:4282332	Pinot blanc	G	G/A	G/G	intron
chr18:4383633	Pinot blanc	G	G/C	G/G	exon

chr18:21103474	Pinot blanc	A	A/C	A/A	intergenic
chr18:28693529	Pinot blanc	G	A/G	G/G	exon
chr19:3781226	Pinot blanc	G	G/A	G/G	intron
chr19:7124786	Pinot blanc	G	G/A	G/G	0.5Kb upstream gene
chr19:17977220	Pinot blanc	G	G/A	G/G	intergenic
chr2:18594509	Pinot blanc	C	C/T	C/C	intron
chr4:3313620	Pinot blanc	T	T/C	T/T	exon
chr4:14367939	Pinot blanc	T	T/C	T/T	intron
chr4:22792206	Pinot blanc	C	C/A	C/C	exon
chr5:1911407	Pinot blanc	T	T/C	T/T	exon
chr5:4492371	Pinot blanc	T	C/T	T/T	intron
chr5:4735149	Pinot blanc	A	A/G	A/A	intron
chr5:15644060	Pinot blanc	C	T/C	C/C	intron
chr5:22745578	Pinot blanc	G	G/A	G/G	0.3Kb dowstream gene
chr5:24754569	Pinot blanc	G	G/A	G/G	3'UTR
chr7:19629893	Pinot blanc	G	G/A	G/G	intergenic
chr8:7696363	Pinot blanc	G	G/G	T/G	0.3Kb upstream gene
chr8:20690777	Pinot blanc	G	A/G	G/G	exon
chrUn:17599622	Pinot blanc	C	T/C	C/C	3Kb dowstream gene
chrUn:40717178	Pinot blanc	C	C/C	C/T	intron
chr15:19571064	Pinot Meunier	A	A/A	A/G	1Kb dowstream gene
chr1:4897066	Pinot Meunier	A	A/A	A/T	exon
chr1:5250416	Pinot Meunier	C	C/C	C/T	intron
chr1:18587330	Pinot Meunier	A	A/A	T/A	3.9Kb dowstream gene
chr1:19490545	Pinot Meunier	G	G/G	G/T	intron
chr10:4630639	Pinot Meunier	C	C/C	G/C	2.9Kb dowstream gene
chr10:10780961	Pinot Meunier	G	G/G	G/C	intergenic
chr10:17545319	Pinot Meunier	c	C/C	T/C	intron
chr11:19240372	Pinot Meunier	T	T/T	T/A	intron
chr12:2388225	Pinot Meunier	G	G/G	C/G	1.4Kb dowstream gene
chr13:2280849	Pinot Meunier	C	C/C	C/T	exon
chr13:18239780	Pinot Meunier	G	G/G	G/A	intron
chr14:4261604	Pinot Meunier	G	G/G	G/A	intron
chr14:6448337	Pinot Meunier	A	A/A	A/G	0.1Kb dowstream gene
chr14:10718368	Pinot Meunier	G	G/G	G/A	exon
chr14:22875026	Pinot Meunier	A	A/A	A/G	intergenic
chr15:6384633	Pinot Meunier	T	T/T	T/A	intron
chr15:11999866	Pinot Meunier	A	A/A	A/G	7.4Kb dowstream gene
chr16:2823996	Pinot Meunier	G	G/G	G/T	0.3Kb dowstream gene
chr17:8703280	Pinot Meunier	A	A/A	A/G	exon
chr18:990499	Pinot Meunier	A	A/A	A/G	exon
chr18:5632015	Pinot Meunier	C	C/C	C/T	2.6Kb upstream gene
chr18:5713155	Pinot Meunier	G	G/C	G/T	1.8Kb dowstream gene

chr18:12490480	Pinot Meunier	T	T/T	T/C	intron
chr18:19156401	Pinot Meunier	G	G/G	G/A	intergenic
chr18:19911528	Pinot Meunier	G	G/G	G/T	intron
chr18:24888322	Pinot Meunier	T	T/T	T/C	0.9Kb dowstream gene
chr18:29047427	Pinot Meunier	C	C/C	T/C	exon
chr19:453661	Pinot Meunier	G	G/G	G/C	0.8Kb dowstream gene
chr19:1365673	Pinot Meunier	A	A/A	A/G	intron
chr19:3359686	Pinot Meunier	C	C/C	C/T	1.2Kb dowstream gene
chr19:23182339	Pinot Meunier	G	G/G	G/A	intron
chr2:8976579	Pinot Meunier	A	A/A	A/G	intron
chr2:18336309	Pinot Meunier	T	T/T	T/G	exon
chr3:642320	Pinot Meunier	T	T/T	T/G	2.2Kb dowstream gene
chr3:7463543	Pinot Meunier	T	T/T	T/C	intergenic
chr4:2484019	Pinot Meunier	C	C/C	T/C	intron
chr4:11645436	Pinot Meunier	G	G/G	G/C	intron
chr5:8535969	Pinot Meunier	C	C/C	C/T	intergenic
chr5:23597312	Pinot Meunier	G	G/G	G/A	2.5Kb upstream gene
chr6:7669464	Pinot Meunier	T	T/T	T/A	0.5Kb upstream gene
chr6:16757889	Pinot Meunier	C	C/C	C/T	intergenic
chr7:6875279	Pinot Meunier	C	C/C	C/A	intron
chr7:9874978	Pinot Meunier	T	T/G	T/T	intergenic
chr8:454197	Pinot Meunier	T	T/T	C/T	exon
chr8:2094998	Pinot Meunier	C	C/C	T/C	intron
chr8:14163314	Pinot Meunier	C	C/C	T/C	exon
chr8:15009839	Pinot Meunier	G	G/G	G/A	exon
chrUn:34747330	Pinot Meunier	C	C/C	C/T	6.8Kb upstream gene
chrUn:39196308	Pinot Meunier	A	A/A	A/T	exon

IV.2.3 Variant detection in 'Sangiovese R24' and 'Sangiovese VCR23'

A total of 6,041,450 variable positions were detected between 'Sangiovese R24' and the reference genome sequence. A total of 6,123,665 variable positions were detected between 'Sangiovese VCR23' and the reference genome sequence. These positions were called using the default parameters of UnifiedGenotyper of the GATK package, version 2.1-13 (McKenna A et al. 2010). The higher number of variant positions with respect to the reference genome than that observed for the 'Pinot' pair might be ascribed to a combination of the more distant genetic relatedness of 'Sangiovese' with the reference genome of PN40024, which share large haplotype

blocks with ‘Pinot’ and the higher average coverage for the ‘Sangiovese’ pair (50X and 53X) than the ‘Pinot’ pair (37X and 35X).

IV.2.4 Validation of the SNP detection pipeline in ‘Sangiovese’

We analysed large chromosomal segments for which ‘Sangiovese’ was (1) homozygous for the reference haplotype, (2) heterozygous with one haplotype identical to the reference, (3) homozygous for a haplotype different from the reference (**Figure 8**). For each region, we calculated false discovery rate (FDR) before and after applying our filtering. The expected false positives are: heterozygous and homozygous SNPs in regions of type 1, homozygous SNPs in regions of type 2, heterozygous SNPs in regions of type 3.

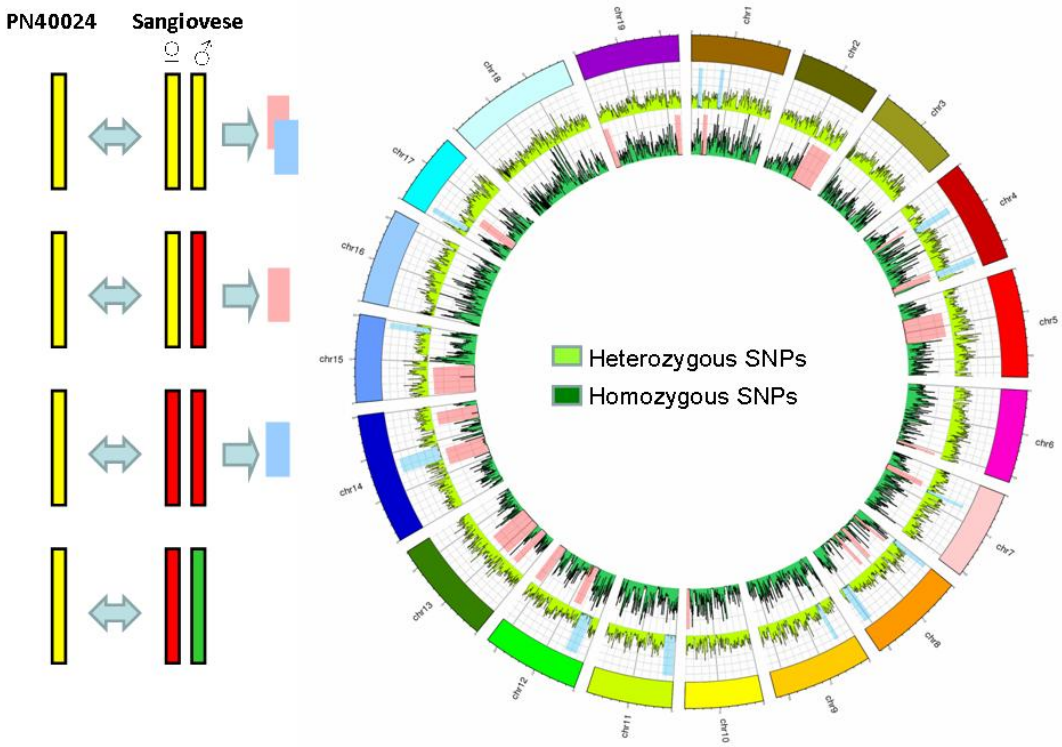


Figure 8 – Circular plots of heterozygous (light green) and homozygous (dark green) SNP density along the nineteen chromosomes of ‘Sangiovese’. Pink boxes indicate regions of type 2, cyan boxes indicate regions of type 3, overlapping of pink and cyan boxes indicate regions of type 1.

Two regions on chr4 and chr14 amounting to approximately 3 million nucleotides were apparently homozygous in 'Sangiovese' for the reference haplotype (**Figure 8**). The borders of the regions were identified by plotting SNP density (**Figure 9**). The region analysed on chr4 was restricted to the interval chr4:21,0..22,3 Mbp. The region analysed on chr14 was restricted to the interval chr14:15,0..17,0 Mbp. In these regions summing up to 3.3 million nucleotides, a total of 4,541 variant positions were called in 'Sangiovese' with respect to the reference genome using the default parameters of UnifiedGenotyper. Of these, 804 variant positions were detected in the interval chr4:21,0..22,3 Mbp (1 variant position every 1,617 nucleotides), the remaining variant positions were detected in the interval chr14:15,0..17,0 Mbp (1 variant position every 535 nucleotides). Based on the calibrated filtering parameters, the number of variant positions was reduced to a total of 90 (). As few as 20, corresponding to 1 false positive SNP every 61,905 nucleotides, remained in the interval chr4:21,0..22,3 Mbp. No homozygous variant SNP was called in this interval. One variant position every 28,986 nucleotides was called in the interval chr14:15,0..17,0 Mbp. The presence of several homozygous variant SNPs with high coverage, high quality scores, and evenly distributed across this interval seems to indicate that this region in 'Sangiovese' has two highly similar haplotypes, each one being slightly different from the haplotype of the reference genome (**Figure 10**) and is therefore not appropriate for the SNP pipeline validation. With an average genome coverage of 53X, FDR in homozygous regions identical to the reference is therefore estimated at 1 false positive SNP every 61,905 nucleotides.

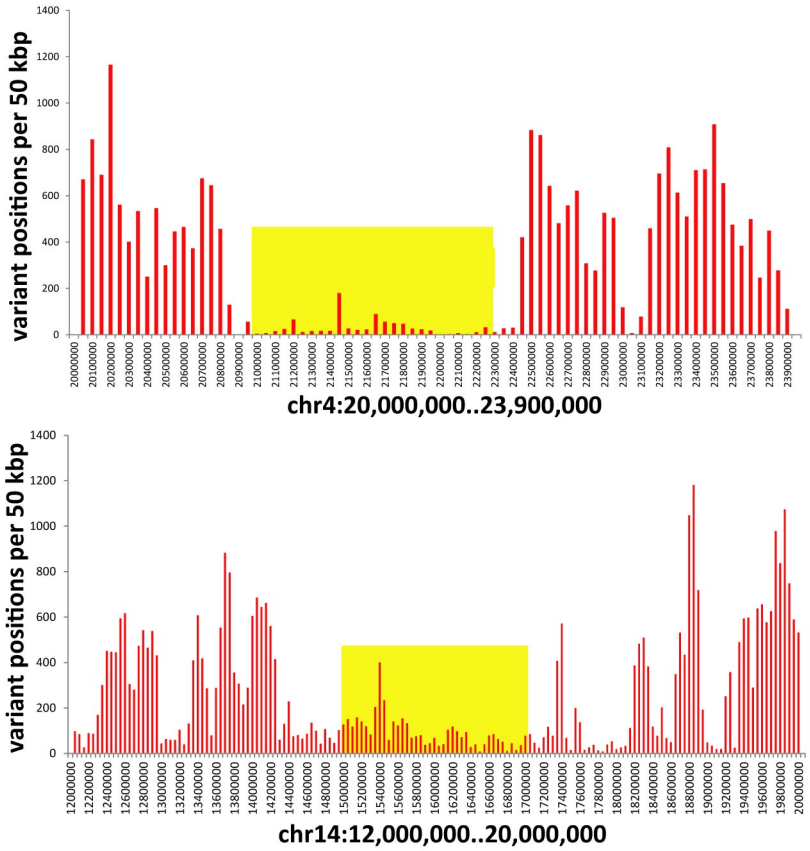


Figure 9 -Density of variant positions between 'Sangiovese' and the reference sequence called using the default parameters of UnifiedGenotyper in two windows on chr4 (above) and chr14 (below). The regions in yellow background were selected for FDR analysis.

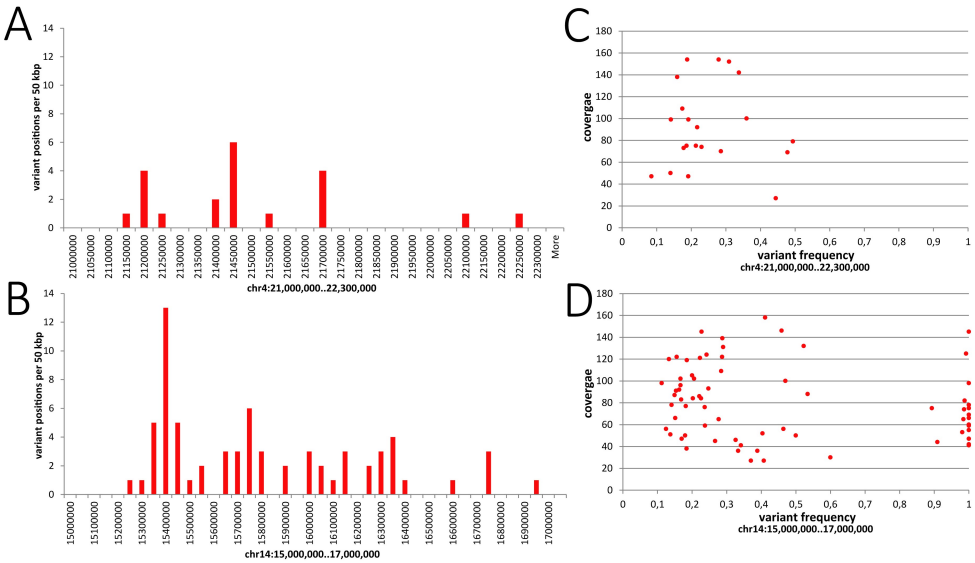


Figure 10 – Density of variant positions between 'Sangiovese' and the reference sequence called using calibrated filtering parameters in two windows on chr4 (A) and chr14 (B). Variant frequency and coverage of each variant position in the two intervals are plotted in panels C and D.

Sixteen chromosomal regions were heterozygous in 'Sangiovese' with one haplotype identical to the reference (**Figure 8**). Except for the large region on chr2, the borders of the other regions were more precisely delimited by plotting SNP density (**Figure 9**). The regions used for FDR analysis in heterozygous regions amounted to 49.2 million nucleotides. A total of 425,496 variant positions were called between 'Sangiovese' and the reference genome sequence, using the default parameters of UnifiedGenotyper, corresponding to 1 variant position every 116 nucleotides. In 364,152 of those variant positions, the variant frequency was comprised between 0.2 and 0.9; in 3,020 variant positions the variant frequency was >0.9; in 58,324 variant positions the variant frequency was <0.2. Based on the calibrated filtering parameters, the number of variant positions was reduced to 110,999, corresponding to 1 variant position every 444 nucleotides. Without applying any specific filtering for allelic frequency, most of the high quality SNPs had a variant frequency comprised between 0.2 and 0.8 (**Figure 12**). Only four SNPs had variant frequency > 0.9 and genotype likelihoods compatible with being called homozygous variant. Based on genotype likelihood values all other SNPs were called heterozygous. FDR of homozygous SNPs in heterozygous regions with one haplotype identical to the reference is 1 false positive homozygous SNP every 12,313,500 nucleotides.

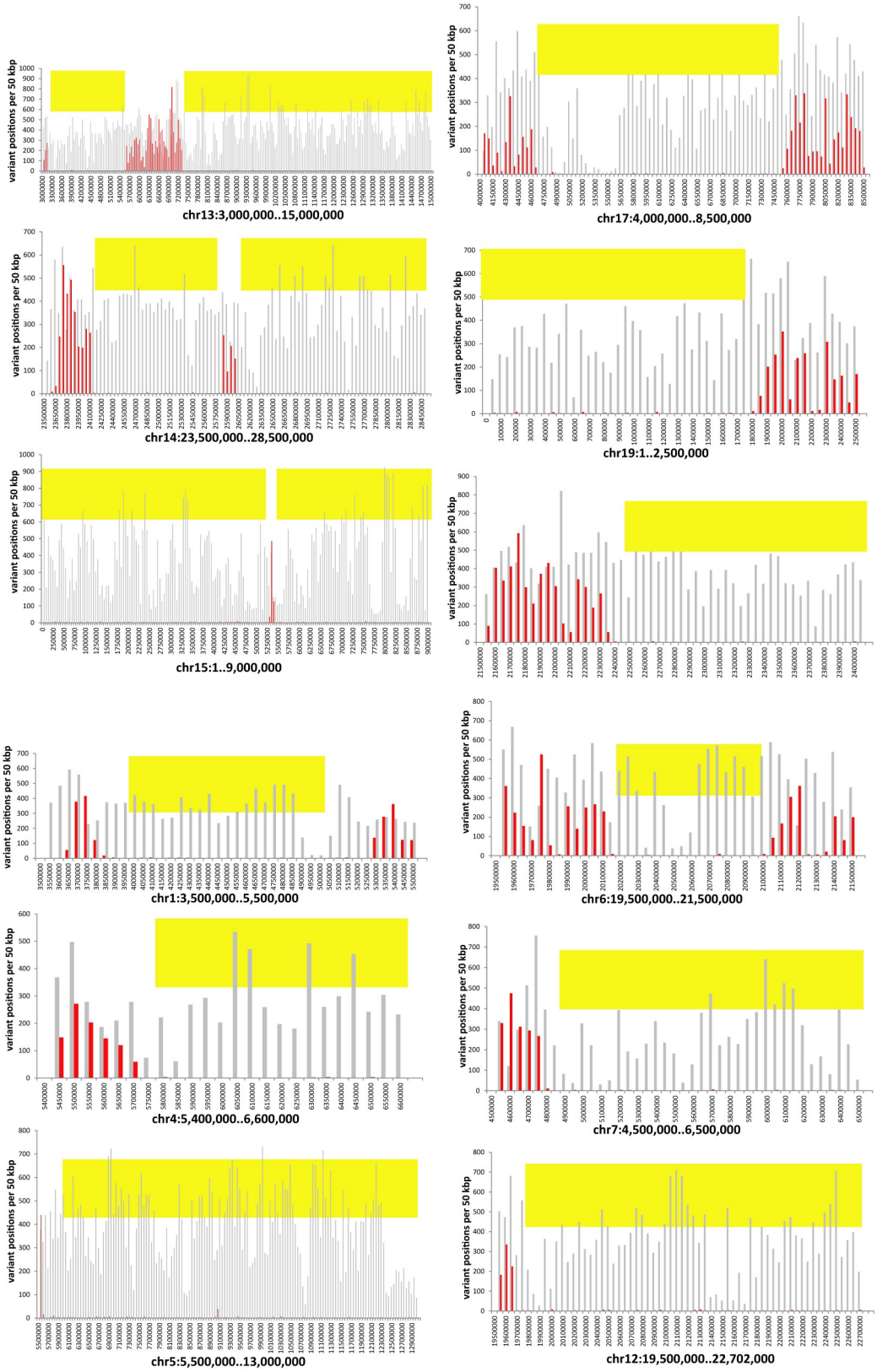


Figure 11 - Density of variant positions between 'Sangiovese' and the reference sequence called using the default parameters of UnifiedGenotyper in 16 chromosomal regions. Heterozygous SNPs with variant frequency comprised between 0.2 and 0.9 are indicated with grey bars,

homozygous SNPs with variant frequency > 0.9 are indicated with red bars. The regions in yellow background were selected for FDR analysis.

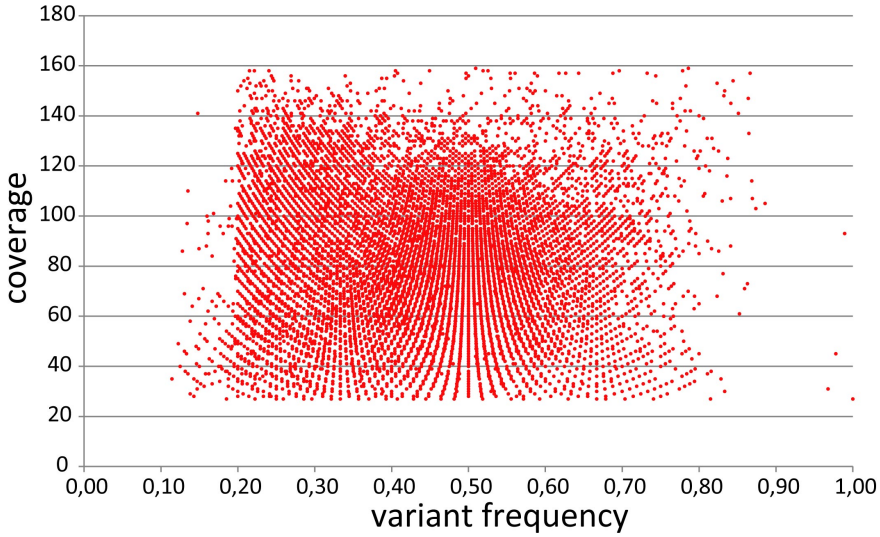


Figure 12 - Variant frequency and coverage of each variant position in heterozygous regions of 'Sangiovese' with one haplotype identical to the reference

Eleven chromosomal regions were homozygous in 'Sangiovese' for a haplotype different from the reference haplotype (**Figure 8**). The borders of these regions were more precisely delimited by plotting SNP density (**Figure 13**). The regions used for FDR analysis in homozygous regions amounted to 16.5 million nucleotides. A total of 172,101 variant positions were called between 'Sangiovese' and the reference genome sequence, using the default parameters of UnifiedGenotyper, corresponding to 1 variant position every 96 nucleotides.

Based on the calibrated filtering parameters, the number of variant positions was reduced to 4,908. Of these, 2,766 variants positions were called homozygous based on genotype likelihoods, corresponding to 1 homozygous SNP every 5,965 nucleotides. Another 2,142 variant positions were called heterozygous based on genotype likelihoods. With an average genome coverage of 53X, the estimated FDR of heterozygous SNPs in homozygous regions for a haplotype different from the reference is 1 false positive SNP every 7,703 nucleotides.

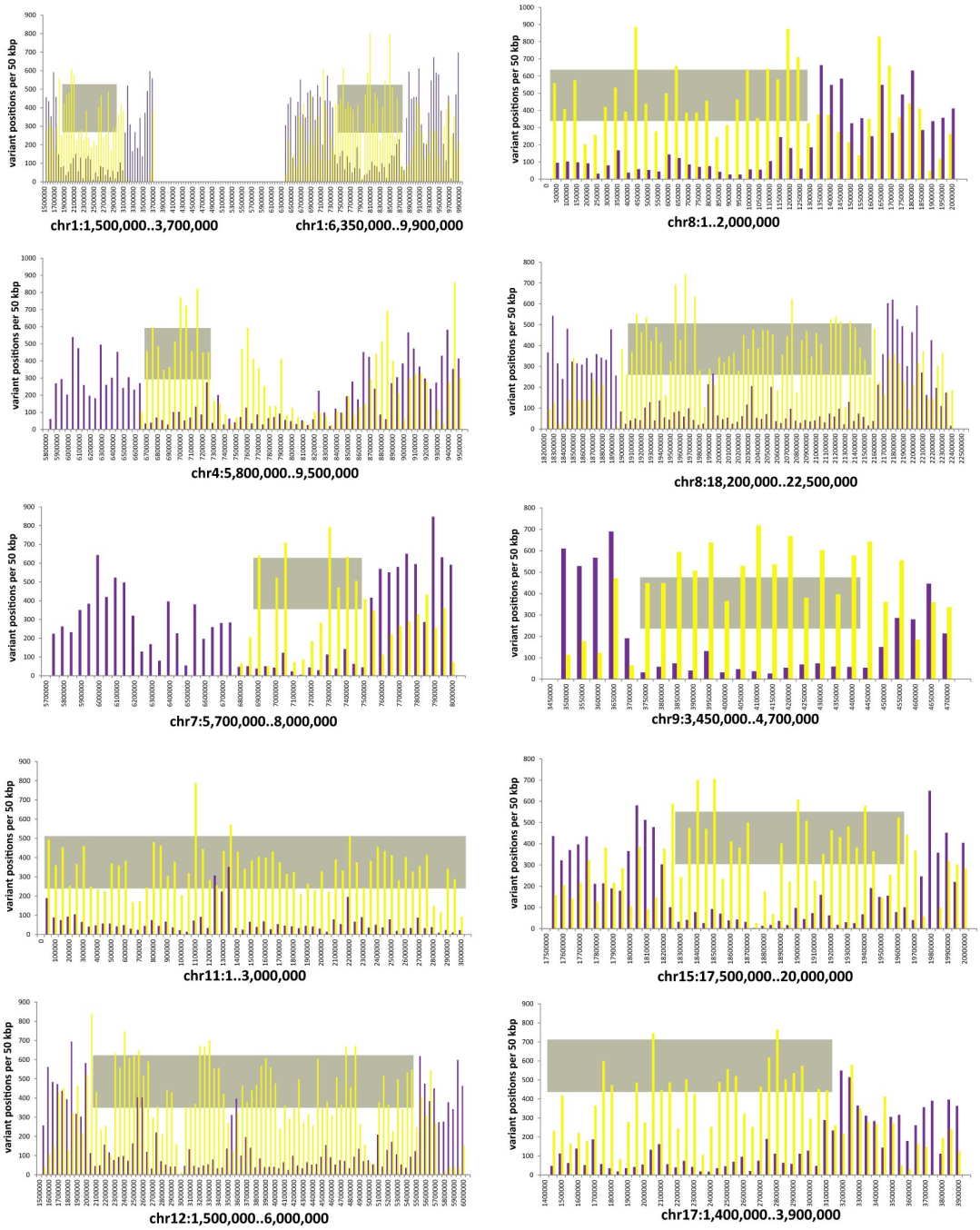


Figure 13 - Density of variant positions between 'Sangiovese' and the reference sequence called using the default parameters of UnifiedGenotyper in 11 chromosomal regions. Heterozygous SNPs with variant frequency comprised between 0.2 and 0.9 are indicated with purple bars, homozygous SNPs with variant frequency > 0.9 are indicated with yellow bars. The regions in grey background were selected for FDR analysis.

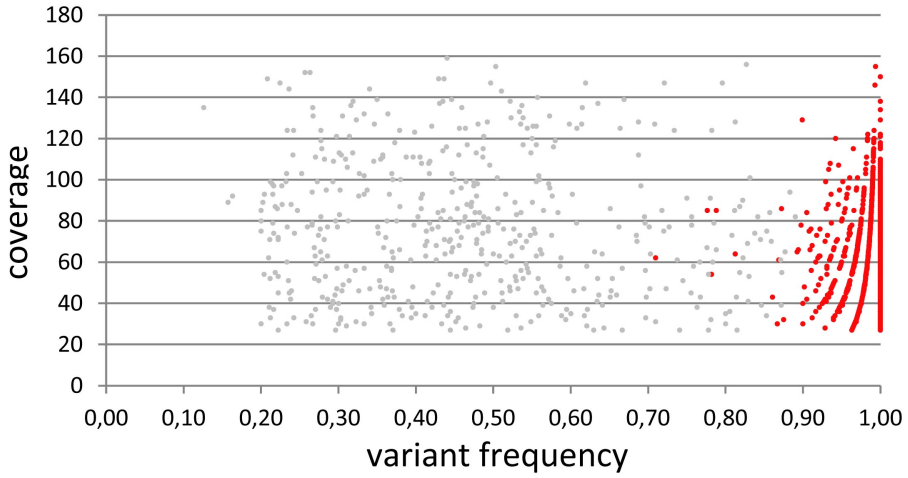


Figure 14 - Variant frequency and coverage of each variant position in homozygous regions of 'Sangiovese' for a haplotype different from the reference. Red dots represent variant positions called homozygous variant based on genotype likelihoods. Grey dots represent variant positions called heterozygous.

The same pipeline with the parameters used for comparing clones within a variety (except for the parameter 'adjacent SNP <100bp'), detected a total of 1,472,993 SNPs between the varieties 'Pinot' and 'Sangiovese' on a whole-genome scale. Of these, 484,620 were shared between the two varieties, 392,987 were unique in 'Pinot Meunier' (351,703 heterozygous and 41,284 homozygous) and 595,144 were unique in 'Sangiovese' (491,864 heterozygous and 103,280 homozygous).

IV.2.4.1 Filtering

Based on the parameters calibrated on 'Pinot Meunier' and 'Pinot blanc', the set of raw SNPs in 'Sangiovese R24' and 'Sangiovese VCR23' was filtered for:

- variable positions shared by both clones
- variable positions in repeated regions, transposable elements and small indel/SSR intervals
- minimum coverage <0.5-fold the average coverage
- maximum coverage >3-fold the average coverage
- GATK Phred-scaled quality score (QUAL) < 100
- GATK FisherStrand (FS) < 0
- GATK Strand Bias (SB) > 0

- Phred-scaled likelihoods (GATK PL) for each of the 'homozygous reference', 'heterozygous', homozygous alternate' possible genotypes: 'true genotype' = 0, others < 50, Σ other <300
- distance from the end of the read for reads with the alternate allele (GATK ReadPosRankSumTest <-2 and >2.5)
- adjacent SNP (< 100bp) without evidence of hemizygosity (no significant difference in genome coverage between individuals)
- minor allele frequency < 0.2

With this filtering, we ended up with 55 variant positions between 'Sangiovese R24' and 'Sangiovese VCR23'. A considerably lower number of variant positions was observed for this pair of clones than for the two 'Pinot' clones. This might be due to the higher average coverage for the 'Sangiovese' pair (50X and 53X) than the 'Pinot' pair (37X and 35X) and/or to a closer relationship among the clones. We inspected every variable position by visualising the aligned reads with Tablet. In 52 cases, we found evidence that the putative variable position was due to a nucleotide variant in phase with neighbouring variants, likely originating from misaligned paralogous sequences, that were completely filtered out in one clone and partially in the other. The visual inspection of these regions confirmed the alignment of non-allelic sequences as the cause for the appearance of these false differences among clones because in most cases three sequence haplotypes were visible in the same region (thanks to the presence of multiple SNPs within the same read), all differing from each other by more than one SNP. This condition was associated with strand bias, since the partially filtered paralogous SNP was present in reads produced from one of the DNA strands. The remaining three SNPs did not show any evidence of bias, and were not located in regions that apparently hampered the correct alignment of mapped reads. Quality scores are given in **Table 3**. Two of these candidate SNPs (chr11:14389639, chr19:22173717) are located in low complexity regions resembling nascent microsatellite repeats (Appendix 1). As for the third SNP (chr13:16,483,189), the mutation has occurred in a region that contains 4 other SNP, all present in the same haplotype, 3 of which are within less than 100 bp from

the new SNP and thus could be localised within the same read. The other haplotype appears to be identical to the reference one in this region. The variant nucleotide in the variant individual is consistently in phase with the reference haplotype only.

The observed variant frequency for all three SNPs is consistent with the range expected for chimerical heterozygous mutations that occurred in a once-homozygous position – once identical to the other clone and to the reference genome – and with the cell layer composition of leaf tissues from which DNA was extracted, that are a mixture of homozygous wild-type and heterozygous mutated genotypes.

Table 3 – Quality scores of the SNPs identified between clones of 'Sangiovese' and comparison with the known SNP at chr1:4,897,066 in 'Pinot Meunier'

position	reference	variant	position quality score	variant clone	reference reads	variant reads	coverage	variant frequency	quality score of the genotype	homozygous reference	heterozygous	homozygous variant	reference reads	variant reads	coverage	variant frequency	quality score of the genotype	homozygous reference	heterozygous	homozygous variant	coverage ratio	ReadPosRankSum
										Phred-scaled likelihoods								Phred-scaled likelihoods				
					R24								VCR23									
chr11:14389639	C	T	184	VCR23	47	0	47	0,00	99	0	135	1661	35	9	44	0,20	99	217	0	1237	1,07	2,333
chr13:16483189	C	T	298	VCR23	55	0	55	0,00	99	0	165	2014	23	11	34	0,32	99	331	0	750	1,62	0,567
chr19:22173717	A	T	250	R24	30	12	42	0,29	99	283	0	1011	42	0	42	0,00	99	0	123	1597	1,00	0,211
					Pinot blanc								Pinot Meunier									
chr1:4897066	A	T	299	Meunier	71	0	71	0,00	99	0	205	2695	48	14	62	0,23	99	335	0	1702	1,15	0,86

IV.2.5 SNP confirmation by capillary sequencing

We experimentally validated these three variants by PCR amplification of the variant position with flanking primers followed by direct Sanger sequencing of the amplicons. We investigated separately genomic DNA extracted from leaves and from pollen tissue of 'Sangiovese R24' and 'Sangiovese VCR23'. Sequence for position chr11:14,389,639 failed the quality filter, whereas putative SNPs in position chr13: 16,483,189 and chr19:22,173,717 turned out to be true somatic variants in clone 'Sangiovese VCR23' and in clone 'Sangiovese R24', respectively (**Table 4**).

Table 4 – Genotype calls from Sanger amplicons sequenced in the experimental validation of unique variant positions selected from the somatic SNP detection pipeline in comparison 'Sangiovese R24' and 'Sangiovese VCR23'. (ND = Not Determinable Data).

			Genotype				
			R24		VCR23		
SNP position	Variant clone	Reference	leaf	pollen	leaf	pollen	
chr11:14389639	VCR23	C	ND	ND	ND	ND	intergenic
chr13:16483189	VCR23	C	C/C	C/C	C/T	T/T	intergenic
chr19:22173717	R24	A	A/T	ND	A/A	ND	intergenic

IV.2.5.1 A detailed analysis of SNP in position chr13:16,483,189

The GATK caller UnifiedGenotyper predicted position in chr13:16,483,189 to be heterozygous C/T in leaf tissue of 'Sangiovese VCR23' whereas 'Sangiovese R24' is homozygous reference C/C. Sanger resequencing of the amplicons across the SNP, confirms a homozygous position C/C for 'Sangiovese R24' in both leaf and pollen tissues. Sanger resequencing in 'Sangiovese VCR23' reveals a different genotype call between tissues (**Figure 15**): the leaf is heterozygous C/T whereas the pollen is homozygous variant T/T.

Since pollen is a germline tissue that originates from the L2 inner layer we expected a homozygous reference C/C genotype for 'Sangiovese VCR23' pollen if the mutation had occurred in the L1 layer or a balanced (50:50) heterozygous C/T

genotype if it had occurred in the L2 layer. A hemizygous region spanning the variant somatic mutation can explain the observation of a homozygous T/T genotype obtained by Sanger sequencing. No structural variant was detected by the pipelines used in this work, therefore we visually inspected the 'Sangiovese VCR23' alignment to find the presence of spanning reads that map at a distance higher than the insert size as evidence of a deletion. No evidence was found, suggesting the need for further investigation.

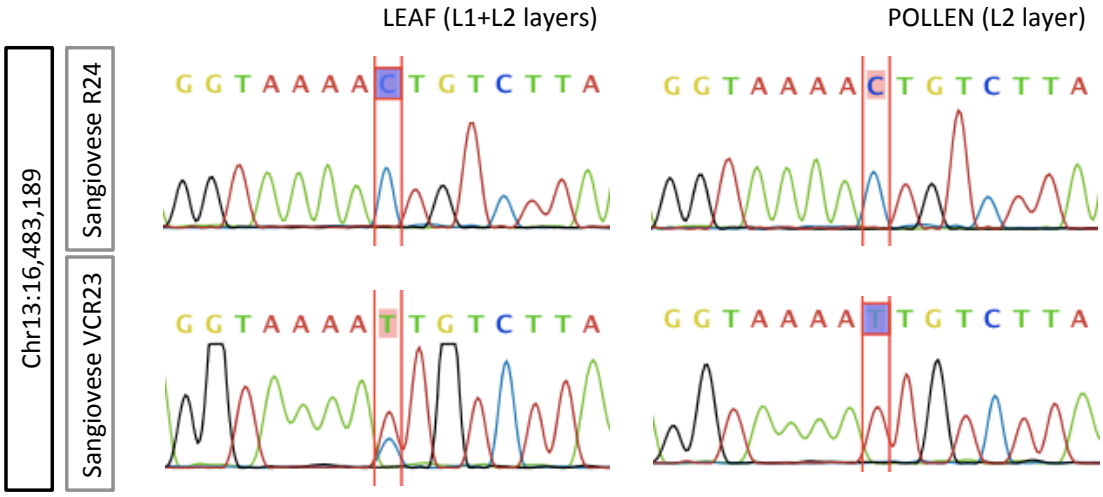


Figure 15 – Comparison of electropherograms obtained by Sanger resequencing of the chr13:16,483,189 region among the studied 'Sangiovese R24' and 'Sangiovese VCR23'. The red line indicates the polymorphic position C/C and T/C in leaf and C/C and T/T in pollen. In particular, 'Sangiovese VCR23' electropherograms shows the heterozygous state in L1+L2 derived tissue (leaf), and the homozygous variant state in pure L2-derived tissue (pollen). 'Sangiovese R24' is fully reference homozygous-state in both L1+L2 derived and pure L2-derived tissues.

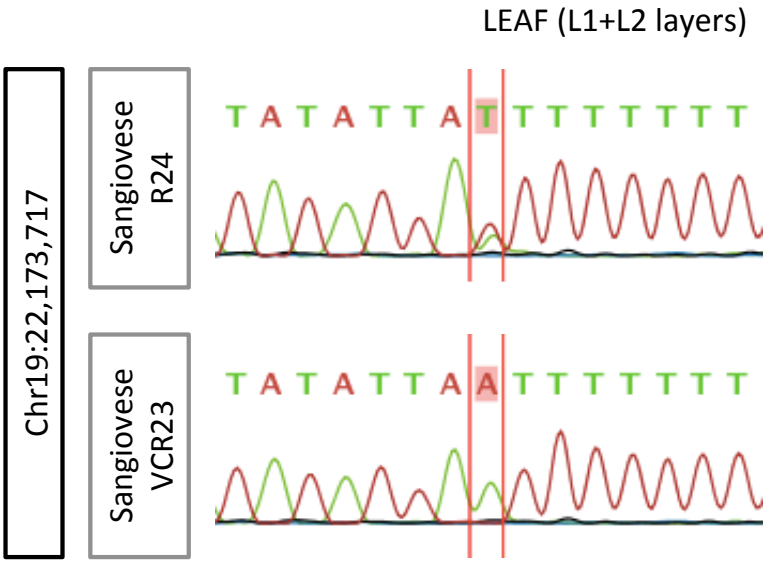


Figure 16 – Comparison of electropherograms obtained by Sanger resequencing of the chr19:22,173,717 region among the studied 'Sangiovese R24' and 'Sangiovese VCR23'. The red line indicates the polymorphic position T/A and A/A in L1+L2 derived tissue (leaf). In particular, 'Sangiovese R24' electropherograms shows the heterozygous state while 'Sangiovese R24' is fully reference homozygous-state. Resequencing of genomic DNA extracted from pollen failed.

IV.2.6 Variant detection in 'Pinot Meunier' and 'Traminer'

The pipeline used to discover variant nucleotides between 'Pinot' clones was applied with the same parameters to analyse single nucleotide differences between 'Pinot Meunier' and 'Traminer'. Since the varieties 'Pinot' and 'Traminer' are connected by a parent-offspring relationship, they share half of the genome in the form of a shared haplotype. This means that the homozygous SNPs among them could be either true positive SNPs in hemizygous genome portions or false positive SNPs in heterozygous regions giving us an estimation of FDR at whole genome scale. A total of 4,749,724 raw variable positions were detected between 'Pinot Meunier' and the reference genome sequence, whereas a total of 5,386,685 raw variable positions were detected between 'Traminer' and the reference genome sequence.

IV.2.6.1 Filtering

Based on the parameters calibrated on 'Pinot Meunier' and 'Pinot blanc', the set of raw SNPs in 'Pinot Meunier' and 'Traminer' was filtered for:

- variable positions shared by both varieties
- variable positions in repeated regions, transposable elements and small indel/SSR intervals
- minimum coverage <0.5-fold the average coverage
- maximum coverage >3-fold the average coverage
- GATK Phred-scaled quality score (QUAL) < 100
- GATK Strand Bias (SB) > 0
- Phred-scaled likelihoods (GATK PL) for each of the 'homozygous reference', 'heterozygous', homozygous alternate' possible genotypes: 'true genotype' = 0, others < 50, Σ other <300
- distance from the end of the read for reads with the alternate allele (GATK ReadPosRankSumTest <-2 and >2.5)
- minor allele frequency < 0.2

With this filtering, we ended up with 556,974 variant positions between 'Pinot Meunier' and 'Traminer'. Among these, 213,013 were unique SNPs for 'Pinot Meunier' and 343,961 were unique SNPs for 'Traminer'. As expected since these two varieties have a shared haplotype, the total number of point mutations is approximately half of the total number of SNPs identified between 'Sangiovese VCR23' and 'Pinot Meunier'. In chromosome 4, where both haplotypes appear to be completely shared between 'Pinot Meunier' and 'Traminer', we identified 995 heterozygous SNPs and 12 homozygous SNPs out of 10,287,597 detectable nucleotides, leading to an estimated FDR of 1 false positive SNP every 10,339 nucleotides for heterozygous SNPs and of 1 every 857,300 nucleotides for homozygous SNPs. However, heterozygous SNPs are not evenly distributed across chromosome 4, but they tend to cluster in specific regions, i.e. the telomeric sequences. Even if 'Pinot Meunier' and 'Traminer' share completely the chromosome 4 and therefore we would expect a low FDR, the FDR values found are not surprising: grapevine genome is highly heterozygous and varieties have genome portions that are not present in the reference sequence of PN40024 and that cause mis-alignments during reference read mapping. Since clones are genetically

identical, portions that mis-align to the reference genome have the same distribution that biases SNP detection in both clones. This is not true when comparing the genomes of two different varieties because the portions that could mis-align are not shared and affect the SNP detection, causing a higher FDR.

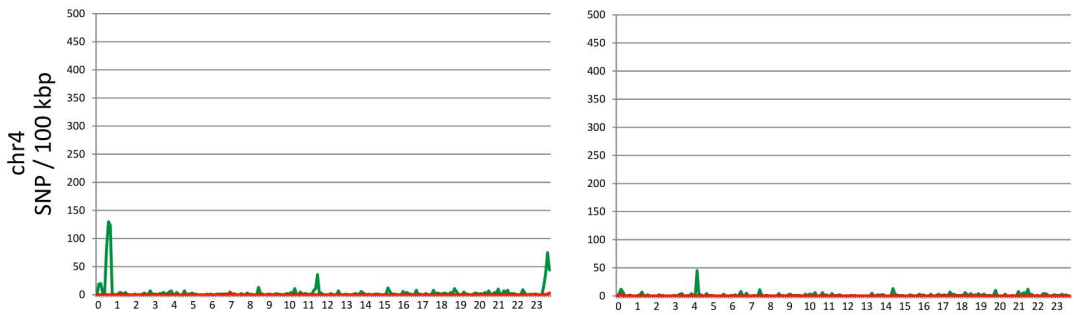


Figure 17 – Distribution of unique SNPs in variety 'Gewürztraminer' and in variety 'Pinot Meunier' resulting from the pairwise comparison along chromosome 4. Heterozygous SNPs are shown in green, while homozygous SNPs are showed in red. Chromosome 4 appears to be completely shared between the two varieties except three portions, two of them corresponding probably to the telomeric regions.

IV.3 Detection of structural variants

In order to discover deletions in the genomes of 'Pinot' and 'Sangiovese' clones, we applied two methods, both focused on mapping sequence reads to the reference genome. The Depth of Coverage approach (DOC) assumes a random distribution (typically Poisson or modified Poisson) in mapping depth and exploits the high coverage of NGS to investigate the divergence from this distribution to the variant event. The number of reads is expected to be proportional to the number of times the region appears in the sample. Methods that use DOC signature must partition the reference into windows so that the coverage depth is consistent within a window but may have a sharp difference between adjacent windows. Due to the fact that the DOC signature is directly related to the absolute number of reads falling within each window and thus to the coverage of the dataset and to the size of the CNV, we use the power of DOC method to detect only large events (> 25 kbp).

For smaller events, we applied the Paired-End Mapping (PEM) approach, which analyses the mapping information of paired-end reads, their discordancy from the expected span size and map strand properties. PEM signatures are more powerful to detect smaller events compared with DOC signatures, but may require higher coverages. Read pairs that map to the reference genome too far apart from expected insert size, define deletions in the sequenced genome, while reads that can only be mapped as singletons and not as pairs may point to novel insertions in the sequenced genome.

IV.3.1 Depth of Coverage analysis

Structural variants between clones due to copy number changes were investigated by analysis of depth of coverage signatures. We set thresholds of 0.8 and -0.8 for the log2 ratio of the number of reads from each clone, normalised to the average genome coverage of that clone, mapped in windows of defined size along the chromosomes and required that at least 10 consecutive windows and 25kb of sequence be involved in the event.

In order to calibrate the parameters for copy number variant (CNV) detection, we used again the comparison between 'Pinot blanc' and 'Pinot Meunier', because they differ by a large heterozygous deletion that was estimated to span ~100-179 kb on chr2:14,149,000..14,250,000 by Vezzulli et al (2012). In the SNP detection of this thesis, we found evidence for loss-of-heterozygosity between the chromosomal positions 14,115,143 and 14,309,249.

Our procedure for structural variant detection identified 71 consecutive windows where 'Pinot blanc' had lower copy number than 'Pinot Meunier' in between the chromosomal positions 14,104,720 and 14,239,561, with an average log2 ratio over the region of 0.8192. The same analysis applied to the comparison between 'Pinot noir' and 'Pinot Meunier' did not reveal any variation, as expected.

In order to validate the presence and the extent of the deletion, we inspected the alignment of all mapped paired-end reads spanning the region of the heterozygous deletion identified by DNAcopy. We found evidence of ten paired-end reads

produced from 400-700 bp inserts that were mapped at larger-than-expected distance on the reference sequence. The spanning paired-ends that mapped most closely to the deletion allowed us to define the borders of the deletion more precisely than the methods of loss-of-heterozygosity and depth-of-coverage. PEM restricted the location of borders of the deletion downstream of chr2:14,105,175 and upstream of chr2:14,258,065, within an expected interval of variation on each side of less than 1 kbp. These evidences provide an estimate for the size of the deletions that is not larger than 153 kbp.

DNA copy identified 34 additional regions in which the copy number was lower in 'Pinot blanc' with respect to 'Pinot Meunier' and were supported by log2 ratios > 0.8. All the detected variants ranged in size between 3.6 and 14.8 kb, and were identified by a number of consecutive windows ranging from 2 to 7 (**Table 5**). Thus, DNAcopy did not detect any other significant large CNV (greater than 25 kb and comprising at least 10 consecutive windows) in 'Pinot blanc' besides the known one in chromosome 2.

With the thresholds of log2 ratio <-0.8 and >25 kb and >10 windows, 21 events were detected where the copy number was lower in 'Pinot Meunier' than in 'Pinot blanc'. These regions ranged in size between 25 and 132 kbp. We visually inspected on the genome browser all these putative events of copy number variation and assessed that all the identified segments correspond to regions with highly repetitive sequence content, resembling centromeric repeats. Thirteen of them were identified on scaffolds not anchored to chromosomes, and several of them have low log2 ratios, supporting the hypothesis that they correspond to regions where alignment of reads is particularly critical.

Table 5 – List of putative CNV events in 'Pinot blanc' and 'Pinot Meunier' identified by analysis of depth of coverage.

deletion in	chromosome	start	end	size	windows	log2 ratio
Pinot blanc	chr2	14104720	14239561	134841	71	0,8192
Pinot Meunier	chr11	15196362	15274360	77998	19	0,8952
Pinot Meunier	chr12_random	1413643	1449891	36248	15	0,9155
Pinot Meunier	chr12_random	1519023	1565642	46619	15	1,1177
Pinot Meunier	chr16	9812070	9885235	73165	18	0,9503
Pinot Meunier	chr2	11909738	11951870	42132	12	1,1676
Pinot Meunier	chr4	12400516	12453144	52628	16	1,1323
Pinot Meunier	chr6	9550801	9587770	36969	14	1,0082
Pinot Meunier	chr9	14912126	14969868	57742	17	0,9794
Pinot Meunier	chrUn	130647	198036	67389	16	1,2306
Pinot Meunier	chrUn	6365127	6436489	71362	17	0,9316
Pinot Meunier	chrUn	13025979	13064052	38073	10	0,9711
Pinot Meunier	chrUn	20251581	20314328	62747	24	0,8943
Pinot Meunier	chrUn	22522535	22568401	45866	18	1,0756
Pinot Meunier	chrUn	24634353	24718330	83977	24	0,9958
Pinot Meunier	chrUn	25204244	25279122	74878	13	0,993
Pinot Meunier	chrUn	25438041	25518061	80020	25	1,2513
Pinot Meunier	chrUn	25743904	25814473	70569	12	1,1597
Pinot Meunier	chrUn	26664961	26797442	132481	38	1,1151
Pinot Meunier	chrUn	28781533	28838478	56945	12	0,9688
Pinot Meunier	chrUn	32924449	32964050	39601	11	1,0942
Pinot Meunier	chrUn	34571056	34595557	24501	10	1,2047

With the same thresholds of log2 ratio > 0.8 or <-0.8 and at least 25 kb and 10 windows, DNAcopy did not detect any large CNV between 'Sangiovese R24' and 'Sangiovese VCR23'.

IV.3.2 Paired-End Mapping

. PEM signatures were exploited to detect structural variation events (both insertions as well as deletions) smaller than 25 kbp.

IV.3.2.1 Deletions

Analysis of paired-end mapping (PEM) data to detect deletions in comparison to the reference sequence was performed using the software BreakDancer. BreakDancer identified 3,086 putative deletions in 'Pinot blanc', 3,991 in 'Pinot gris', 4,381 in 'Pinot Meunier', and 1,171 in 'Pinot noir' when each one of them was compared to the PN40024 sequence. Since we expect somatic mutations to correspond to novel

events, we compared the deletions detected in each of the 'Pinot' clones against those identified in a group of 20 varieties of *Vitis vinifera* analysed with the same pipeline and excluded those that were also found in at least another variety from further analyses. After such a filtering step, eleven putative deletions in 'Pinot blanc', nineteen in 'Pinot gris', fifteen in 'Pinot Meunier', and five in 'Pinot noir' were identified as unique to each clone.

BreakDancer identified 5,193 putative deletions in 'Sangiovese R24' and 3,596 in 'Sangiovese VCR23'. Of these, 1,643 putative deletions were unique to 'Sangiovese R24' and 46 putative deletions to 'Sangiovese VCR23'. This asymmetry in the number of candidate variants identified in the two clones is likely due to the difference in the number of mapped paired-ends in the two clones. Despite a lower coverage in terms of total mapped reads (50X versus 53X), 'Sangiovese R24' had a higher proportion of reads mapped in pairs, which are the only type of reads that are useful for this analysis.

Given the high number of candidate regions, we focused on those that were unique to 'Sangiovese VCR23'. Of the initial 46 putative deletions, 39 were also found in the same set of 20 varieties of Vitis vinifera analysed with the same pipeline. These deletions are typical of the variety, and present in other varieties that share with 'Sangiovese' the same haplotype in the region of the deletion. These deletions were missed by the pipeline in the other clone. The remaining 7 putative events of deletions identified in 'Sangiovese VCR23' are not shared with other varieties of Vitis vinifera and deserve further validation.

IV.3.2.2 Insertions

Insertions were investigated with a proprietary pipeline developed at the Institute of Applied Genomics. The procedure detects the presence in a single genome region of two adjacent groups of unpaired reads in opposite orientation, orphaned by the nonalignment of their mates, which are then assembled into sequence contigs and compared to databases of transposable elements.

This pipeline identified 1,791 putative insertions in ‘Pinot blanc’, and 3,071 in ‘Pinot Meunier’., when each of them was compared to the PN40024 sequence. Of these, 146 putative events are shared by all four clones of ‘Pinot’. In a similar way to what occurred with PEM, this asymmetry in the number of candidate regions identified is likely due to the difference in the number of mapped paired-ends in the different clones, which is the only set of useful reads for this analysis. Since we expect somatic mutations to correspond to novel events, we compared the insertions detected in each of the two ‘Pinot’ clones against those identified in a group of 20 varieties of *Vitis vinifera* analysed with the same pipeline and in ‘Pinot gris’ and ‘Pinot noir’ and excluded those that were also found in at least another genotype from further analyses. After such a filtering step, 119 putative insertions in ‘Pinot blanc’ and 160 in ‘Pinot Meunier’ were identified as unique to each clone. The annotation of these regions is underway.

The pipeline identified 1,324 putative events of insertion shared by the two clones of ‘Sangiovese’; 4,366 putative events of insertion only in ‘Sangiovese R24’ and 531 only in ‘Sangiovese VCR23’. In a similar way to what occurred with PEM, this asymmetry in the number of candidate regions identified in the two clones is likely due to the difference in the number of mapped paired-ends in the two clones, which is the only set of useful reads for this analysis. Given the high number of candidate regions, we focused on ‘Sangiovese VCR23’. Of the initial 531 putative insertions, 357 were also found in a set of 20 varieties of *Vitis vinifera* analysed with the same pipeline. These insertions are typical of the variety, and present in other varieties that share with ‘Sangiovese’ the same haplotype in the region of the insertion. These insertions were missed by the pipeline in the other clone. These insertions are true positives at varietal level, but false positives in the comparison between clones. The remaining 174 putative events of insertion identified in ‘Sangiovese VCR23’ are not shared with other varieties of *Vitis vinifera* and deserve further attention. They might be either transpositions that occurred in a rare haplotype that is present in ‘Sangiovese’ and not shared by any of the other varieties investigated – and not identified by the pipeline in ‘Sangiovese R24’ – or

more recent events of transposition that occurred specifically in one clone. Additional analyses on this topic are underway.

Table 4 – List of putative insertion events detected as unique in 'Pinot blanc', 'Pinot Meunier', and events detected as unique for 'Sangiovese R24 ' and 'Sangiovese VCR23' through the analysis of paired end mapping.

chr	start	end			
	Insertion		**length**	**Variant clone**	**content**
chr1	20021850	20021858	9	Pinot blanc	Ty1-copia
chr3	5149952	5150079	53	Pinot blanc	Ty3-gypsy
chr3	4652170	4652227	12	Pinot Meunier	Ty1-copia
chr3	5537013	5537286	58	Pinot Meunier	Ty1-copia
chr8	8409504	8409565	5	Pinot blanc	Ty1-copia
chr12	18113514	18113566	128	Pinot Meunier	NA
chr14	3321677	3321741	116	Pinot blanc	Ty3-gypsy
chr14	22419360	22419373	147	Pinot blanc	Ty1-copia
chr17	2057407	2057418	67	Pinot Meunier	Retrovirus
chr1	903759	903811	53	Sangiovese R24	Ty3-gypsy
chr12	8519666	8519674	9	Sangiovese R24	NA
chr19	21087537	21087715	179	Sangiovese R24	Ty1-copia

IV.4 Global transcriptional changes

IV.4.1 Filtering and alignment of Illumina reads

A total of 30 libraries corresponding to 30 independent RNA samples and biological replicates were produced (**Table 6**). The raw reads were processed for adapter removal, quality trimming and filtering for contaminants. Post-processed reads were aligned to the grapevine transcriptome and to the reference genome using TopHat version 2.0.5 (Trapnell et al. 2012). In the first phase of alignment, reads were mapped to the grapevine transcriptome defined by the version V1 of gene

annotation. Unmapped reads were aligned against the reference genome allowing for gaps in order to permit reads to span exons and discover novel transcripts (**Table 5**). As a result of this process, for leaf transcriptome, 30,759 genes were tested for differential expression in the comparison of 'Pinot' clones, and 30,006 genes in the comparison of 'Sangiovese' clones. For 'Sangiovese' berry transcriptome, 30,648 genes were tested for differential expression at stage 1 (2 weeks after berry set) and 30,283 genes at stage 2 (inception of ripening).

Table 6 – List of libraries for RNA-seq. Number of high quality reads and fraction of aligned reads

tissue	clone	biological replicate	trimmed and filtered reads	mapped reads	% mapped reads
leaf	**Pinot blanc**	1	27695104	22400818	80,9
leaf	**Pinot blanc**	2	26869276	22503180	83,8
leaf	**Pinot blanc**	3	33529848	21856711	65,2
leaf	**Pinot gris**	1	31499239	26237334	83,3
leaf	**Pinot gris**	2	32832191	27342702	83,3
leaf	**Pinot gris**	3	32155696	26587646	82,7
leaf	**Pinot Meunier**	1	25681174	21630894	84,2
leaf	**Pinot Meunier**	2	34801597	28624979	82,3
leaf	**Pinot Meunier**	3	37999112	31745756	83,5
leaf	**Pinot noir**	1	35545423	29053001	81,7
leaf	**Pinot noir**	2	51079364	42338949	82,9
leaf	**Pinot noir**	3	42690794	34491920	80,8
leaf	**Sangiovese R24**	1	34178282	29302470	85,7
leaf	**Sangiovese R24**	2	24077066	20442743	84,9
leaf	**Sangiovese R24**	3	27049960	22954613	84,9
leaf	**Sangiovese VCR23**	1	38066030	32397577	85,1
leaf	**Sangiovese VCR23**	2	17468554	14882097	85,2
leaf	**Sangiovese VCR23**	3	31460168	27224055	86,5
berry stage1	**Sangiovese R24**	1	67431840	55791338	82,7
berry stage1	**Sangiovese R24**	2	50208878	41509057	82,7
berry stage1	**Sangiovese R24**	3	54833017	45629228	83,2
berry stage1	**Sangiovese VCR23**	1	69228909	57876256	83,6
berry stage1	**Sangiovese VCR23**	2	35880120	29813908	83,1
berry stage1	**Sangiovese VCR23**	3	88132256	75079590	85,2
berry stage2	**Sangiovese R24**	1	49829685	41058675	82,4
berry stage2	**Sangiovese R24**	2	57000116	46812915	82,1
berry stage2	**Sangiovese R24**	3	57665663	47009333	81,5
berry stage2	**Sangiovese VCR23**	1	55723003	44910236	80,6
berry stage2	**Sangiovese VCR23**	2	65363824	52819704	80,8
berry stage2	**Sangiovese VCR23**	3	86057616	71775300	83,4

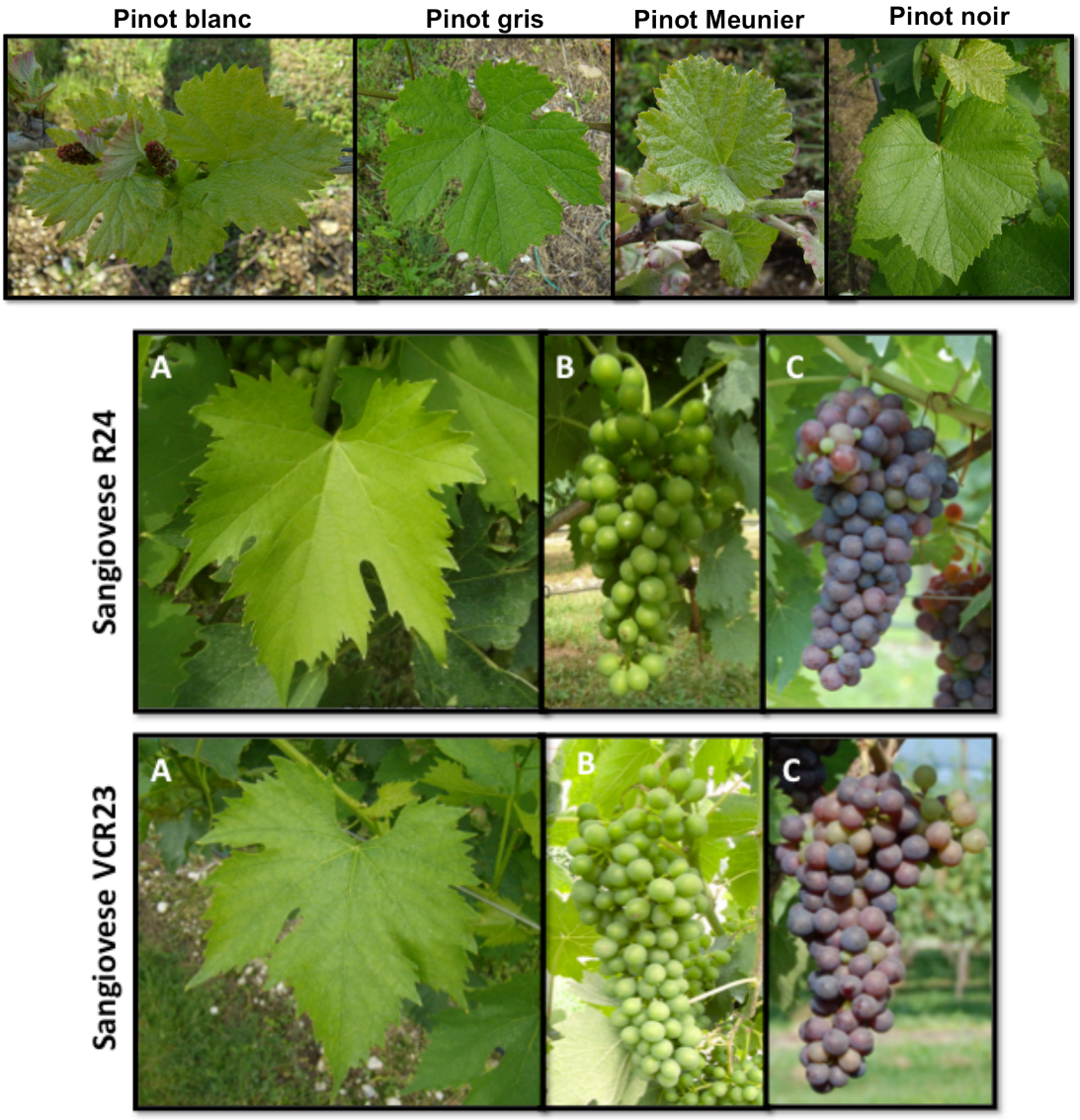


Figure 18 - Tissues harvested for RNAseq analysis of 'Pinot' clones and 'Sangiovese' clones. Three biological replicates were sampled from three vegetatively propagated plants per clone planted along the raw in the vineyard net to each other. Each biological replicate was separately processed during the all procedure. (A) Each leaf tissue replicate consists of a mixture of the most distal leaves along the shoot. Berries were sampled before ripening (B) and at the inception of ripening (C) – see Materials&Methods.

IV.4.2 Differentially expressed genes in leaves of 'Pinot' clones

A number of genes variable between ~200 and 2,200 were differentially expressed in each pairwise comparison among the four clones of 'Pinot' when using a threshold of log2 fold change > 0.5 (**Table 7**) and a statistical significance of P<0.05.

Table 7 – Number of differentially expressed genes between 'Pinot' clones in leaf tissues

	Pinot blanc	Pinot gris	Pinot Meunier	Pinot noir
Pinot blanc	-	823	2232	825
Pinot gris	-	-	821	219
Pinot Meunier	-	-	-	1598
Pinot noir	-	-	-	-

In order to strengthen the significance of the observed differences, we performed multiple-comparisons, figuring out three possible scenarios based on the model of evolution among clones (**Figure 19**). The first is the most likely one.

1. 'Pinot blanc', 'Pinot gris' and 'Pinot Meunier' all arose from independent mutations of 'Pinot noir'. Under this hypothesis, the mutation in one clone would cause a significant down- or up-regulation of a gene or a cluster of genes in that clone compared to all others, while all other clones should not display statistically significant differential expression between them. A total of 627 genes fall in this category. Of these, 376 were differentially expressed in 'Pinot Meunier', 206 in 'Pinot blanc', 36 in 'Pinot noir', and 9 in 'Pinot gris'. The complete list of these genes is given in Supplementary File 'Expression analysis.xls – spreadsheet 'Pinot clones''.
2. One clone arose from a mutation that occurred in an already mutated clone (sequential mutations). Under this hypothesis, a gene or a cluster of genes would display statistically significant down- or up-regulation in two pairwise comparisons, while the other two pairwise comparisons should not display statistically significant differential expression. A total of 200 genes fall in this category.
3. One clone displays significant over-expression of a gene or a cluster of genes compared to two other clones, and the same clone displays significant down-regulation for the same gene or a cluster of genes in comparison to the fourth clone. A total of 88 genes displayed this expression profile.

In addition to these categories of expression profiles, genes that appeared as differentially expressed in a single pairwise comparison or in all six pairwise comparisons were classified as false positives. Approximately 70% of the

differentially expressed genes reported in **Table 7** fell in this category and were ignored for subsequent analyses.

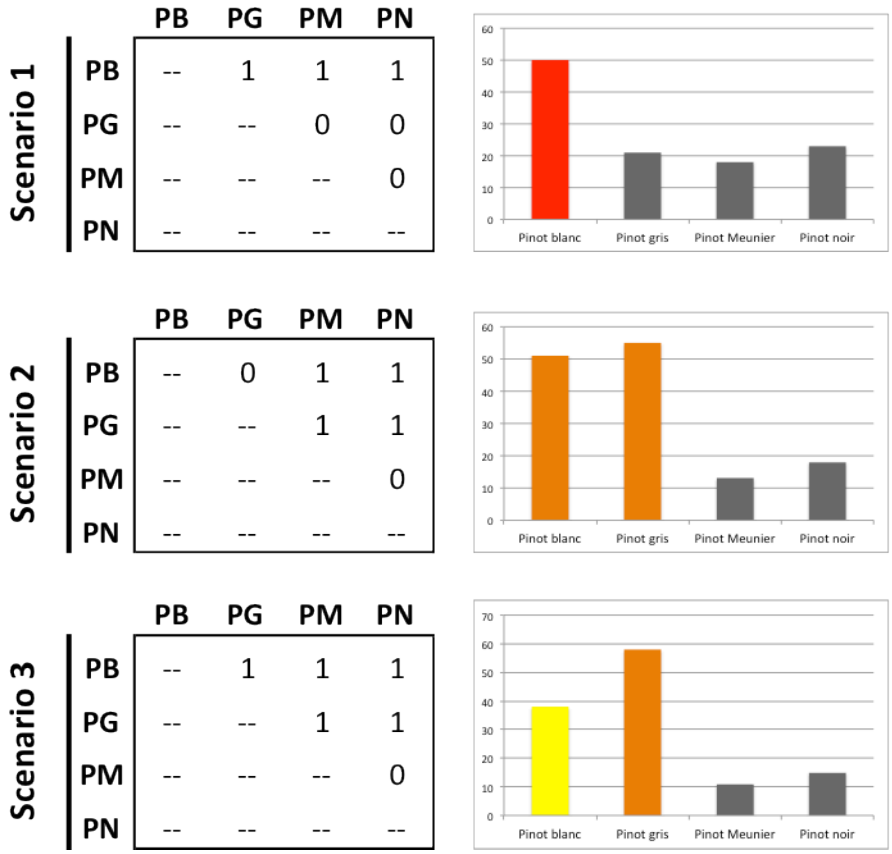


Figure 19 – Expected cases of differential gene expression under different hypothesis of independent (scenario 1) or sequential (scenario 2) somatic mutations. The value 1 and 0 indicate the expected significant (1) and not significant (0) differences of gene expression in pairwise comparisons. The scenario 3 is expected to appear if two successive somatic mutations occur to the same coding region (for instance, insertion and partial excision of a transposable element). In the tables: PB='Pinot blanc', PG='Pinot gris', PM='Pinot Meunier', PN='Pinot noir'.

In order to summarise by category the genes that are differentially expressed among clones, we assigned each gene to a functional category of Gene Ontology using BlastX and Blast2GO (Conesa et al., 2005) searches and to metabolic pathways using MapMan. With this information, differentially expressed genes were classified according to:

- molecular function
- biological process

- metabolic pathway

Most of the differentially expressed genes were classified in the molecular function of catalytic activity and binding (**Table 8**) and in the biological processes of metabolic (**Table 9**) and cellular processes (**Table 10**).

Table 8 – Number of differentially expressed genes between 'Pinot' clones, identified according to scenario 1, grouped by molecular function.

	Pinot blanc		Pinot gris		Pinot Meunier		Pinot noir	
	genes	**%**	**genes**	**%**	**genes**	**%**	**genes**	**%**
catalytic activity	101	45%	4	36%	204	50%	13	33%
binding	95	43%	4	36%	158	39%	24	62%
transporter activity	14	6%	2	18%	21	5%	2	5%
transcription factor activity	5	2%	-		17	4%	-	
enzyme regulator activity	5	2%	-		4	1%	-	
receptor activity	2	1%	1	9%	2	1%	-	
molecular transducer activity	1	0%	-		3	0%	-	

Table 9 – Number of differentially expressed genes between 'Pinot' clones, grouped by biological process.

	Pinot blanc		Pinot gris		Pinot Meunier		Pinot noir	
	genes	**%**	**genes**	**%**	**genes**	**%**	**genes**	**%**
metabolic process	110	25%	4	17%	218	27%	26	45%
cellular process	105	24%	5	22%	177	22%	22	38%
response to stimulus	60	14%	3	13%	133	17%	3	5%
localization	34	8%	2	9%	51	6%	7	12%
biological regulation	26	6%	1	4%	31	4%	-	
developmental process	20	5%	3	13%	49	6%	-	
multicellular organismal process	19	4%	-		47	6%	-	
signaling	17	4%	1	4%	29	4%	-	
cellular component organization	14	3%	1	4%	24	3%	-	
reproduction	12	3%	2	9%	19	2%	-	
death	8	2%	-		7	1%	-	
growth	5	1%	1	4%	15	2%	-	
multi-organism process	3	1%	-		5	1%	-	

The differentially expressed genes in pairwise comparisons between 'Pinot' clones were displayed onto diagrams of metabolic pathways. Compared to the ancestral

state of ‘Pinot noir’, genes associated with secondary metabolism, light reaction and Calvin cycle were down-regulated in leaves of ‘Pinot blanc’ (**Figure 20**); genes associated with secondary metabolism were up-regulated in ‘Pinot gris’ (**Figure 21**); flavonoid genes were down-regulated and phenylpropanoid genes were up-regulated in ‘Pinot Meunier’ (**Figure 22**).

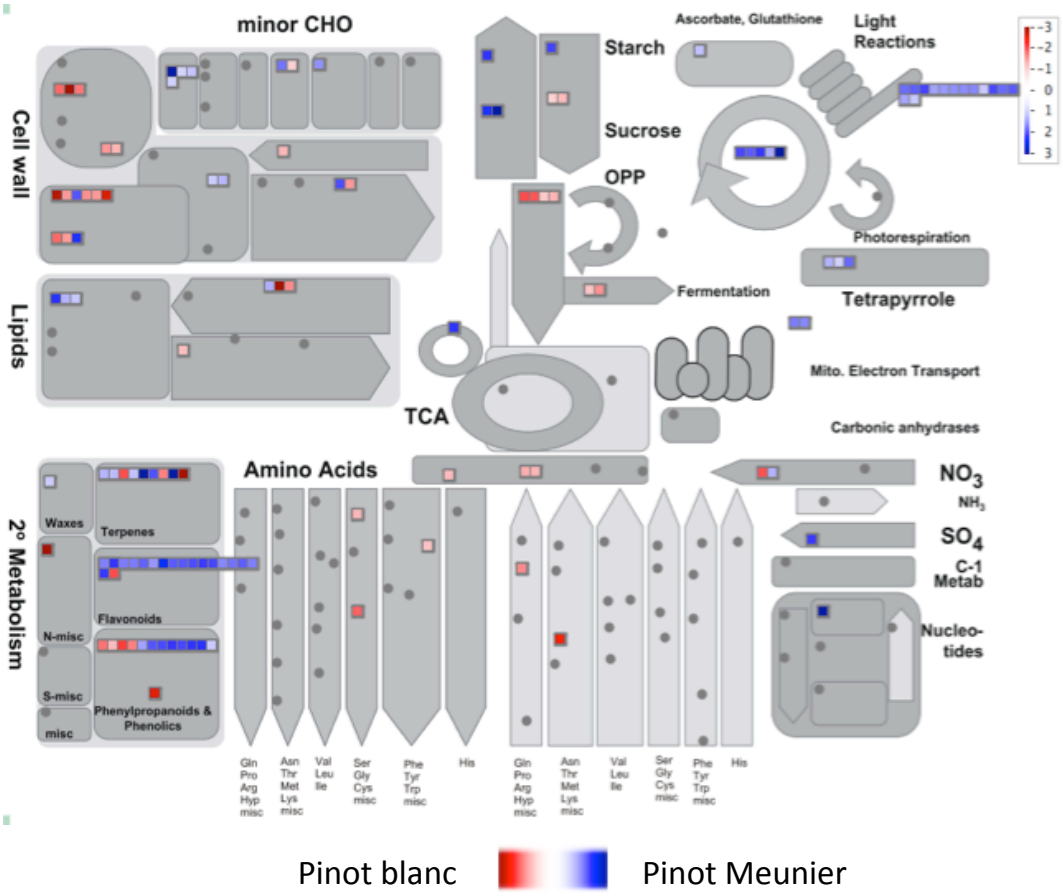


Figure 20 – Metabolism overview of the genes differentially expressed between ‘Pinot blanc’ and ‘Pinot noir’ leaves. The heat map indicates over-expression in ‘Pinot blanc’ (red) or ‘Pinot noir’ (blue).

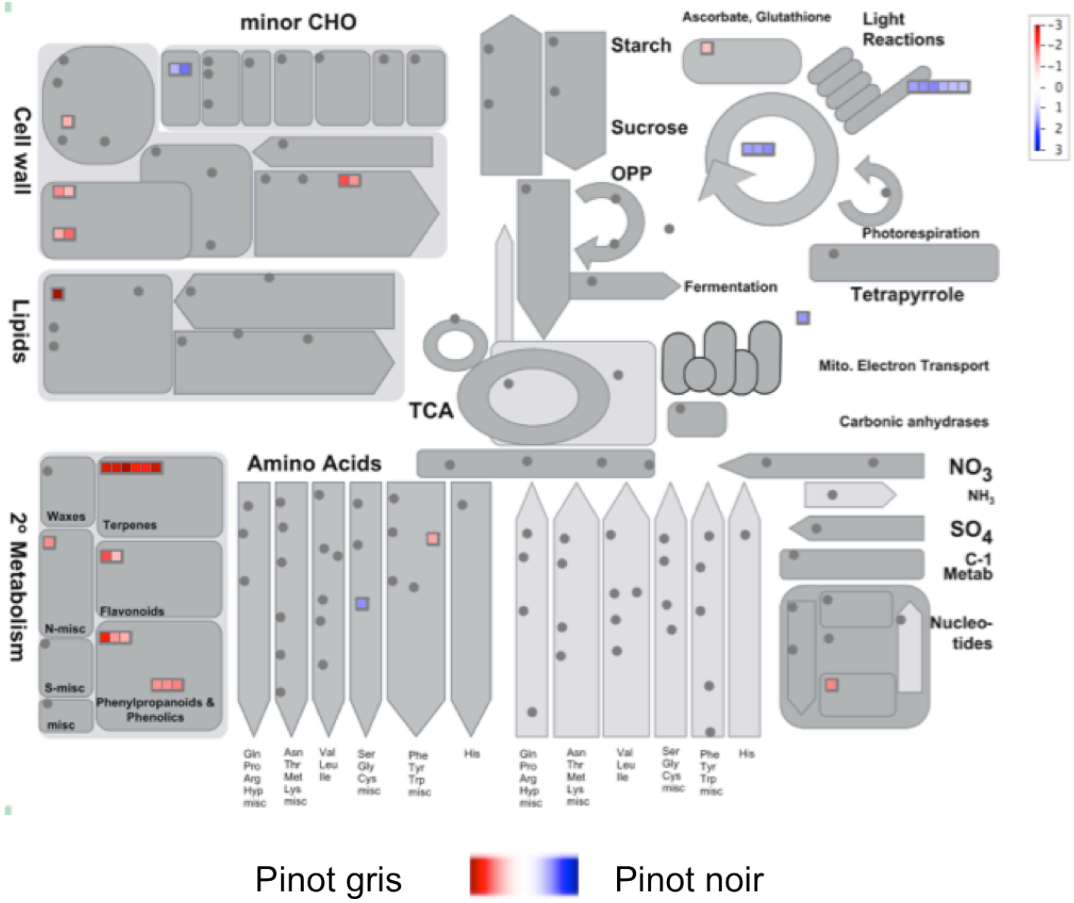


Figure 21 – Metabolism overview of the genes differentially expressed between 'Pinot gris' and 'Pinot noir' leaves. The heat map indicates over-expression in 'Pinot gris' (red) or 'Pinot noir' (blue).

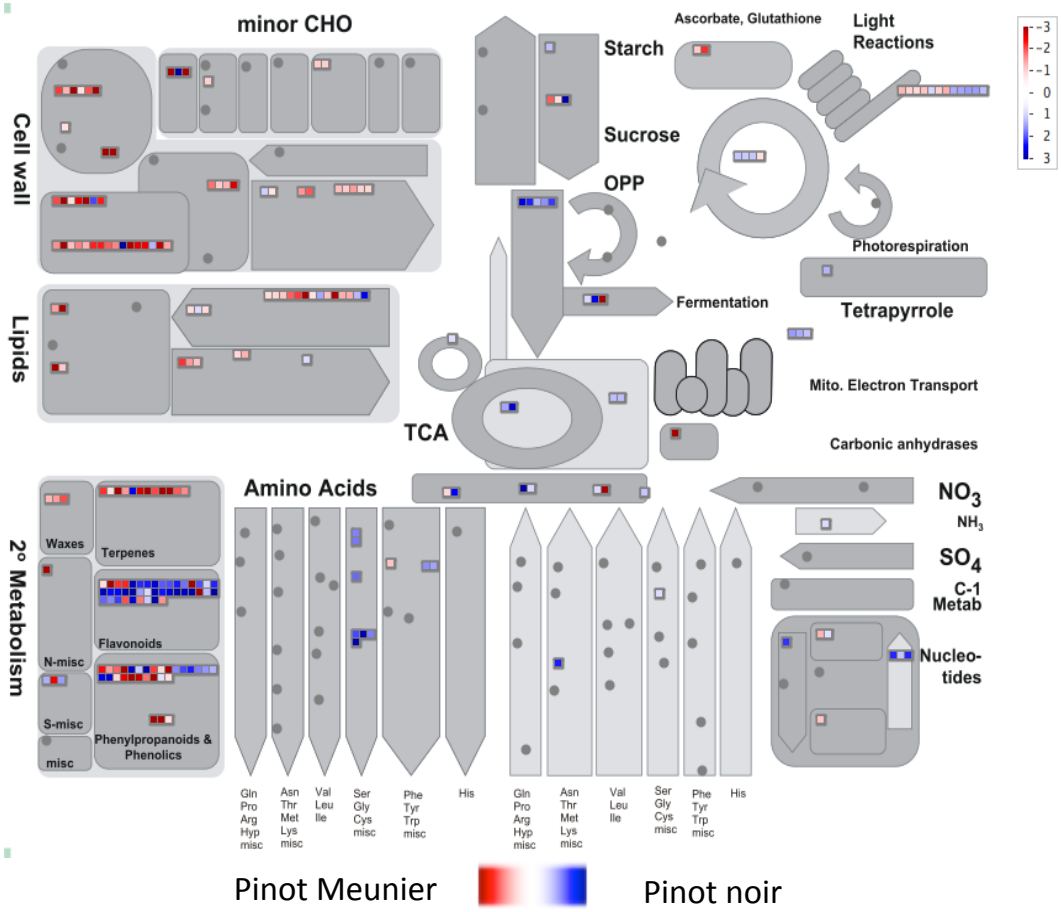


Figure 22 – Metabolism overview of the genes differentially expressed between 'Pinot Meunier' and 'Pinot noir' leaves. The heat map indicates over-expression in 'Pinot Meunier' (red) or 'Pinot noir' (blue).

Focusing on genes associated with secondary metabolism (**Figure 22**), genes involved in the synthesis of chalcones, lignin, and lignans were consistently down-regulated in leaves of 'Pinot blanc'. Genes involved in the synthesis of terpenoids were highly up-regulated in 'Pinot gris', along with a few other involved in phenylpropanoid and simple phenols biosynthesis. Most of the genes involved in the synthesis of terpenoids were up-regulated and most of the genes involved in the synthesis of chalcones were down-regulated in 'Meunier'.

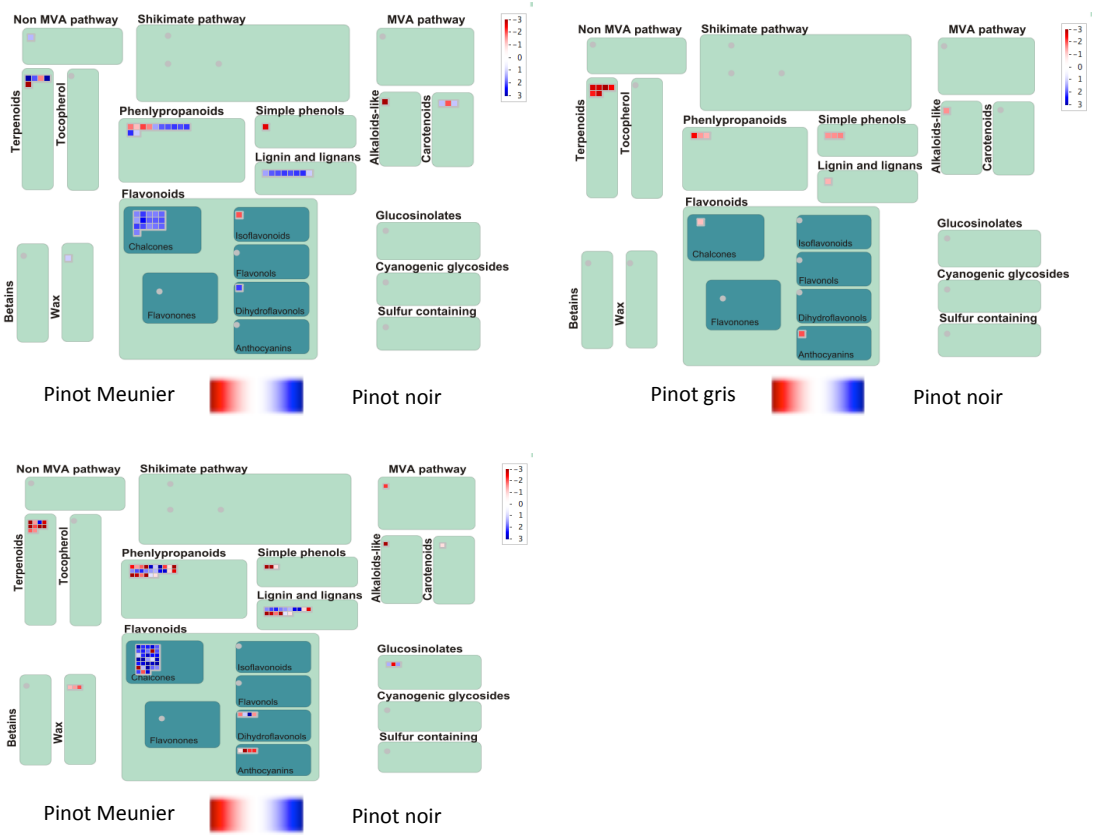


Figure 23 – Overview of the genes involved in secondary metabolism and differentially expressed among 'Pinot' clones. The heat map indicates over-expression in the somatic variant (red) or in the wild-type 'Pinot noir' (blue).

IV.4.3 Differentially expressed genes in leaves and berries of 'Sangiovese' clones

A total of 36, 704, and 103 genes were differentially expressed between 'Sangiovese R24' and 'Sangiovese VCR23' with the threshold of log2 fold change > 0.5 in leaves, berries before ripening and berries at the inception of ripening, respectively (**Table 10**). The complete list of these genes is given in Supplementary File 'Expression analysis.xls – spreadsheet 'Sangiovese clones''. With respect to the pairwise comparisons of 'Pinot' clones, the leaf transcriptome between 'Sangiovese R24' and 'Sangiovese VCR23' displayed much fewer significant changes. Compared to the leaf transcriptome, berry transcriptome was much more variable between 'Sangiovese R24' and 'Sangiovese VCR23', in particular at the first sampling date 2 weeks after berry set, well ahead of the inception of ripening.

Table 10 – Number of differentially expressed genes between 'Sangiovese' clones.

tissue	differentially expressed genes	up-regulated in 'Sangiovese R24'	up-regulated in 'Sangiovese VCR23'
leaf	36	14	22
berry stage1	704	381	323
berry stage2	103	88	14

In terms of magnitude of the differences, the genes differentially expressed in the leaf showed higher values of fold change between the two clones compared to the genes differentially expressed in the berry (**Figure 24**). In berries at the stage 2 (inception of ripening), five genes were over-expressed in 'Sangiovese R24' with a log2 fold-change higher than 4, compared to 'Sangiovese VCR23' (**Figure 24**). These

genes are predicted to encode proteins with similarity to a Fringe-like glucosyltransferase, a zinc finger-homeodomain (ZF-HD) homeobox protein, a lipid-transfer protein, two subtilisin-like proteases.

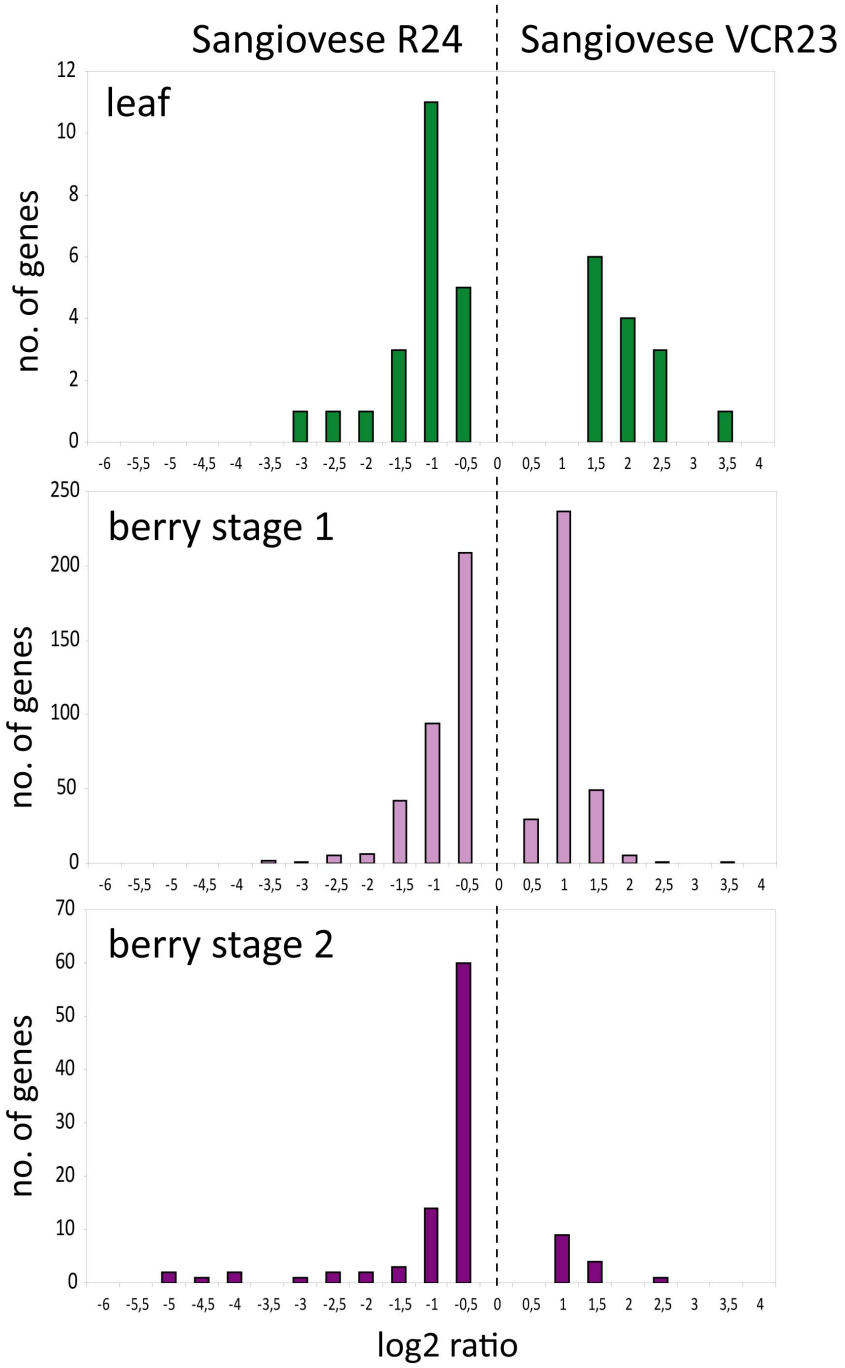


Figure 24 – Distribution of log2 fold change in expression levels of 843 differentially expressed genes between 'Sangiovese' clones

When the differentially expressed genes between ‘Sangiovese R24’ and ‘Sangiovese VCR23’ were displayed onto diagrams of pathways of secondary metabolism, a group of genes involved in chalcone biosynthesis were consistently up-regulated in berries of ‘Sangiovese R24’ collected two weeks after fruit set. Another group of genes for simple phenols, phenylpropanoid, lignin and lignin biosynthesis were slightly down-regulated in berries of ‘Sangiovese R24’ collected at the same stage of sampling. In berries collected after the inception of ripening, a few genes involved in secondary metabolism were differentially expressed between the two clones, except for three genes involved in the synthesis of dihydroflavonols that were more expressed in ‘Sangiovese R24’.

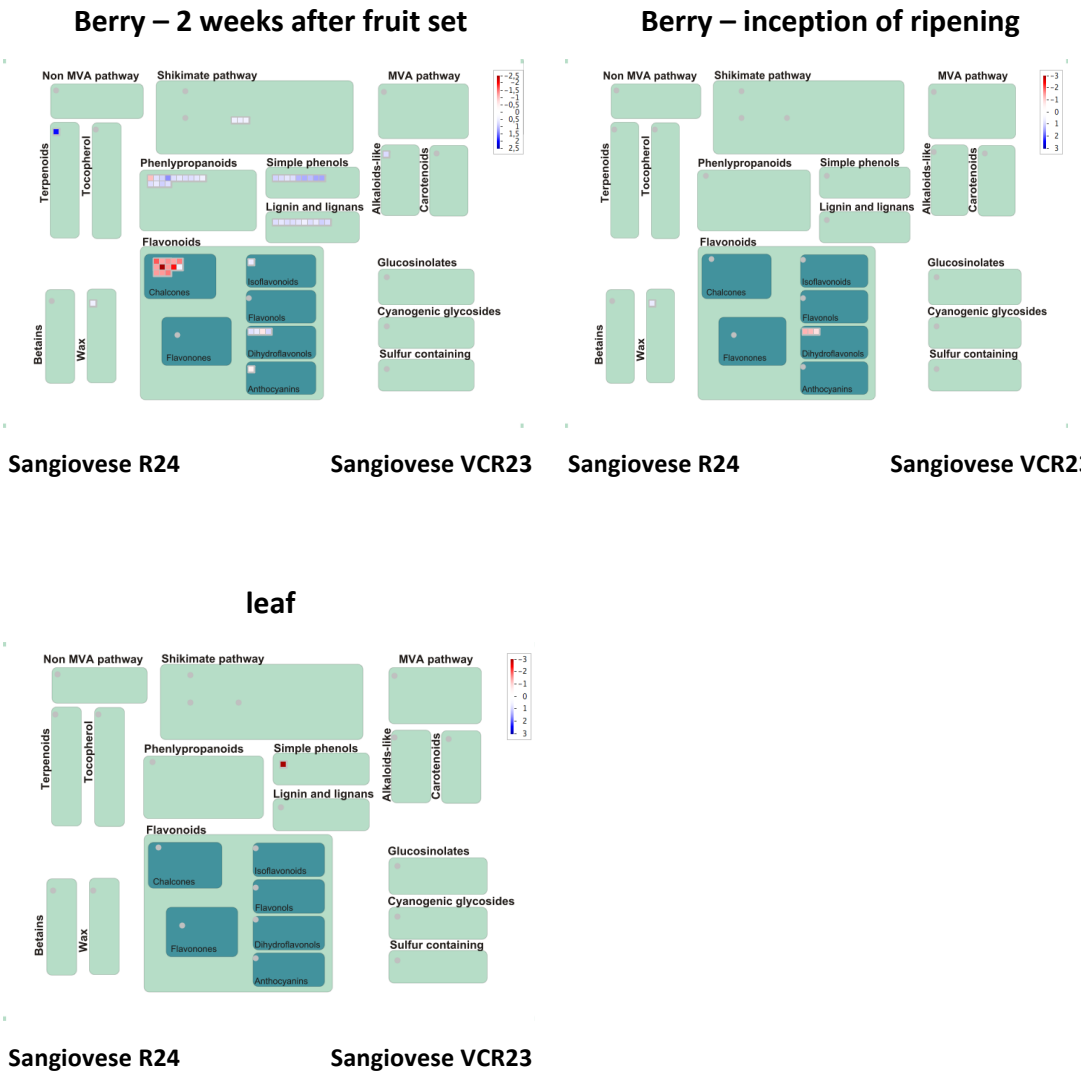


Figure 25 – Overview of the genes involved in secondary metabolism and differentially expressed between 'Sangiovese R24' and 'Sangiovese VCR23' in berries at two developmental stages and in leaves. The heat map indicates over-expression in 'Sangiovese R24' (red) or 'Sangiovese VCR23' (blue).

IV.4.3.1 Differential expression versus chromosomal location of SNP detected between 'Sangiovese' clones

The chromosomal location of the differentially expressed genes between 'Sangiovese' clones was compared with the location of the SNP variants identified by DNA sequencing, in order to check if any SNP located up-stream of a coding sequence might have affected gene expression. The only case of physical vicinity between SNPs and differentially expressed genes was identified for the gene VIT_19s0027g01890 that encodes a vacuolar amino acid transporter and is located

at chr19:22,163,197..22,166,698, some 10 kbp downstream of the SNP chr19:22,173,717, in the absence of any other intervening coding sequence. The VIT_19s0027g01890 vacuolar amino acid transporter gene was found to be 1.6 fold more expressed in ‘Sangiovese VCR23’ than in ‘Sangiovese R24’ in berries collected 2 weeks after fruit set.

V Discussion

Especially for fruit trees, such as Citrus species (Moore et al. 2001) or grapes (This et al. 2006), where vegetative reproduction is used to propagate new interesting phenotypes, somatic variation is very important for genetic improvement and represents a valuable source of heritable mutation. Clonal somatic variation may involve a variety of events such activation of transposable elements, variation in sequence copy number, alteration in chromosome number and structure, gene mutation, somatic crossing-over, sister chromatid exchange, deletion and change in methylation pattern. Somatic variation can be detected using a wide range of techniques having their own strengths and limitations. Until now, previous studies in clonal diversity in grapevine mainly focused on SSRs and AFLP markers (Riaz *et al.* 2002, Hocquigny *et al.* 2004). Although SSRs markers can be very helpful to distinguish grapevine cultivars, Imazio et al. showed that SSRs were not a powerful tool for clonal distinction of V.vinifera 'Traminer' (Imazio *et al.* 2002) and Schellenbaum *et al.* demonstrated that microsatellites are not helpful either for the detection of clones for a specific grapevine cultivar or for the detection of somaclonal variation arising from tissue culture in V.vinifera (Schellenbaum *et al.* 2008). Moreover, AFLP markers offer the possibility to distinguish somatic mutations, but enabled only limited identification of clones (Stenkamp et al 2009).

In spite of the relevance of somatic variation, of the large number of available assays and of the high number of recent publications, little is known about the genetic and epigenetic mechanisms causing this variation. For this reason, a genome-wide approach was chosen to study the somatic variation among four ‘Pinot’ clones and among two ‘Sangiovese’ clones.

V.1 SNP variants in somatic mutants

In clonal identification a goal has to be fulfilled: markers should provide a high discrimination power and they have to be stable, meaning that they produce consistent and repeatable results. Although single nucleotide mutations have a quite low mutation rate when compared with microsatellites, they have been identified to be potential markers to study clonal diversity (Cabezas et al 2011, Carrier et al 2012). Moreover SNPs are highly reproducible and more stable than microsatellites.

In the present study, by comparing 'Pinot blanc' and 'Pinor Meunier', we assessed that NGS provides enough power to detect DNA point mutations in somatic clones, even in the most critical case of chimerical mutations. We set up a bioinformatics pipeline that limits the number of false positive SNPs to a point that currently represents an acceptable compromise between stringency and looseness. Higher stringency filters out chromosomal positions where low coverage (sampling error) and paralogy lead to false positive heterozygous variants with one of the two alleles having a frequency much lower than 0.50, but it would also cause false negative (true variants) chimerical heterozygous variants to be removed. We demonstrated that a known somatic variant present in chimerical homozygous/heterozygous state in leaf tissues of 'Pinot Meunier' (Franks *et al.* 2002) is detected by NGS with 48 reads supporting the evidence for the presence of the wild-type allele and 14 reads supporting the evidence for the presence of the mutated allele. In our libraries, the coverage in that position was sufficiently high to select the variant with high quality scores. The allele ratio of 0.23 for the known mutation falls short of the range of variation expected for heterozygous SNPs with high coverage, but it is hardly distinguishable from false positive SNPs with similar allele ratio even after the Sanger resequencing of the PCR amplified flanking region. By visual inspection of the alignments of many uncertain cases and by Sanger evaluation, most of the false positive SNPs were due to paralogous variants that were incompletely filtered out in either individual under comparison, in regions with critical coverage. In the case of 'Sangiovese' the bioinformatics pipeline returned a list of 55 variant positions above

the imposed quality thresholds, 52 of which were subsequently invalidated by visual inspection of the read alignments. In a similar way as it occurred between 'Pinot' clones, false positives between clones of 'Sangiovese' were due to incomplete filtering in either clone of positions corresponding to mis-aligned reads. The depth of genome coverage seems to be a key factor for reducing the number of these positive variants.

V.2 Structural variants in somatic mutants

The detection of structural variants is of particular importance to characterize somatic clones since in tissue cultured cells, the predominant type of variation is the result of changes in chromosome structure, such a chromosome breakage and in some instances, subsequent exchange or fusion of fragments (Lee and Phillips 1988).

By using a depth of coverage (DOC) approach to investigate copy number variants larger than 25 kbp, we were able to detect only the known somatic deletion in 'Pinot blanc' on chr2:14,149,000..14,250,000 when compared to 'Pinot Meunier' (Vezzulli *et al.*, 2012). Although DOC signature shows 21 putative somatic events (> 25 kbp) where 'Pinot Meunier' has lower copy number in comparison to 'Pinot blanc', the localization in the genome around putative centromeric region, and the high degree of repetitiveness of the involved sequences, suggest that these regions are particular critical for the alignment step. The comparison of PEM signature results revealed 11 putative deletions smaller than 25kbp in 'Pinot blanc', 19 in 'Pinot gris', 15 in 'Pinot Meunier', and 5 in 'Pinot noir' as unique to each clone and not shared with set of 20 varieties of *Vitis vinifera* analysed with the same pipeline.

Since the coverage of uniquely paired sequences that were mapped onto the reference genome is two-fold higher in 'Sangiovese R24' than in the other clone, we focused the analysis of structural variants on those that were detected only in 'Sangiovese VCR23'. Indeed, since the analysis involves the detection of rare somatic events, the discrepancy of coverage between samples causes a higher

number of false positive detection in ‘Sangiovese R24’. All putative large deletions identified by the DOC signature between ‘Sangiovese’ clones turned out to be false positives not showing any evidence of somatic deletion. On the other hand, PEM algorithm identified seven putative deletions events in ‘Sangiovese VCR23’ that are clone-specific and not shared with the set of deletions in 20 varieties of *Vitis vinifera* analysed with the same pipeline. The insertion detection pipeline identified 174 putative events unique in ‘Sangiovese VCR23’ and not shared nor with ‘Sangiovese R24’ nor with other varieties of *Vitis vinifera.* Since the pipeline for the detection of insertions using paired end mapping suffers from a higher false negative rate than that for the detection of deletions (Pinosio S. 2012), these candidates require careful validation. Although we identified small-acquired deletions in ‘Pinot’ and ‘Sangiovese’ clones by using a paired end strategy, we were not able to characterize to the base level structural variation events. Despite the whole genome power detection and the advantages of NGS, none of the described computational approaches to discover structural variation using sequencing is comprehensive. Even if different algorithms are applied to the same genome mapping samples, a significant fraction of the validated variants remains unique to a particular approach. Even though depth of coverage approach was accurate in predicting deletions events, the breakpoint resolution is poor and still needs to be investigated. Instead, Paired End Mapping algorithm revealed to be powerful, but conversely it is dependent on the insert size of its library. We encountered difficulties in repetitive regions, which are indeed more variable and the sensitivity for detecting variation for most of the sequenced-based computational methods is low.

From this study we learned that accurate identification of genetic variation depends both on coverage depth and especially on the alignment of sequenced reads to the correct genomic location. A significant amount of information is lost as a consequence of the resistance of structural variants to proper assembly, misinterpretation of hemizygosity as homozygosity, or because of the

characteristics of grapevine reference genome. Therefore, some regions have to be excluded from the investigation of somatic point or structural mutations since in those portions reads are ambiguously placed or there are an unexpectedly high or low numbers of aligned reads. Furthermore, sequencing and resequencing of the grapevine genome revealed a highly repetitive genome (Jaillon *et al.,* 2007) reducing the ability to map reads uniquely, and showed a high level of heterozygosity preventing the use of a higher stringency during the mapping of the NGS sequences onto the reference. Although we masked the reference genome in repetitive regions and limited ourselves to detecting variation outside of them, the SNP detection among 'Pinot Meunier' and 'Traminer' and the false positive SNPs revealed by PCR amplification and resequencing even after all the filtering steps, showed that there is still a need to define a precise genomic portion that can be interrogated as 'accessible genome' to reduce false-positive detections.

V.3 Transcriptional differences among clones

More than 30,000 genes were expressed in all clones of both varieties. The vast majority of the predicted genes in the grapevine genome was transcribed at detectable levels in all organs and stages of development investigated (leaves, berries before ripening, and berries at the inception of ripening).

Under the same experimental conditions, leaf transcriptomes were much more variable in pairwise comparisons between 'Pinot' clones than between the pair of clones investigated in 'Sangiovese'. This parallels the lower level of DNA sequence differences identified among the two 'Sangiovese' clones than among the Pinot clones and would suggest that the level of genetic identity between 'Sangiovese R24' and 'Sangiovese VCR23' is higher than that existing among 'Pinot' clones. In the leaf transcriptomes of 'Pinot' clones, we found that specific gene categories, mostly associated with secondary metabolism, were differentially expressed among clones. This finding represents an extension of what was already known in berries, where the organ-specific expression of MybA genes – contained in the chromosomal

region involved in the somatic deletions of 'Pinot blanc' and 'Pinot gris' – impairs anthocyanin biosynthesis and causes perturbation to the whole flavonoid pathway (Hocquigny *et al.* 2004; Walker *et al.* 2006; Yakushiji *et al.* 2006). We showed that the transcriptome associated with secondary metabolism is also altered in the leaves.

Between the clones of 'Sangiovese', the widest differentiation in terms of global transcriptome was detected in berries collected two weeks after fruit set. This developmental stage precedes the inception of sugar accumulation and fruit softening, which are associated with ripening. Green and hard berries two weeks after fruit set are in the stage of maximum rate of accumulation of hydroxycinnamic acids and other simple phenols in the flesh, and flavonoids in skins and seeds. In this stage, the synthesis of condensed tannins in skins and seeds is particularly intense. Hydroxycinnamic acids and other simple phenols are synthesised at early steps of the phenylpropanoid pathway, while condensed tannins are formed from phenylpropanoid precursors that enter the flavonoid pathway through their conversion into chalcones. Among the genes differentially expressed between 'Sangiovese' clones in berries collected two weeks after fruit set, we found several genes of the phenylpropanoid and simple phenols pathways expressed at higher levels in 'Sangiovese VCR23' and several genes of the flavonoid pathway leading to the synthesis of chalcones expressed at higher level in 'Sangiovese R24'. In berries collected after the inception of veraison, a few genes involved in secondary metabolism were differentially expressed between the two clones, except for a few genes involved in the synthesis of dihydroflavonols, which were more highly expressed in 'Sangiovese R24'. Dihydroflavonols are important intermediates in the synthesis of anthocyanins and proanthocyanins (or condensed tannins).

VI Conclusions

In this thesis we described the identification of molecular polymorphisms generated and sometime selected during vegetative propagation at the whole-genome scale. The analysis revealed the potential of massively parallel sequencing technology either for DNA resequencing or for transcriptome sequencing for the investigation of differences in the genomes of somatic clones. In fact we were able to distinguish four 'Pinot' clones and two 'Sangiovese' clones even if the two 'Sangiovese' clones chosen for this analysis are evolutionarily very close. For 'Sangiovese' genomes the identified somatic mutations appear to be extremely rare and we have not detected evidence of a relationship between the confirmed mutations and the phenotypic differences observed in the field. Mutations that have occurred in the regions that have not been accessible to us may be those relevant for the phenotypic differences but we cannot rule out the possibility that such differences may result not from genetic but from epigenetic variation. We have been able to detect with a single base resolution a large heterozygous deletion that is responsible for the appearance of white berries in 'Pinot Blanc' and also the already described point mutation in the *GAI1* gene of 'Pinot Meunier'. The importance of the newly defined somatic mutations between 'Pinot' clones for the phenotypic differences among them is not yet known and needs more investigation. Additional analyses are now underway in order validate structural variations and to confirm our results on other clones.

VII Materials and Methods

VII.1 Plant material

Leaf tissues for DNA and RNA extraction were sampled from mother stocks of certified clones, held at the experimental station CASA40 of the Vivai Cooperativi Rauscedo, Italy. The certified clones were 'Pinot Blanc R5', 'Pinot gris R6', 'Pinot Meunier', 'Pinot noir VCR18', 'Sangiovese R24', and 'Sangiovese VCR23'. Leaves were ground in liquid nitrogen before proceeding with DNA/RNA extraction.

Pollen grains for DNA extraction and SNP validation assay were collected from the same plants selected for sampling 'Sangiovese R24' and 'Sangiovese VCR23' leaf green material and ground with tissuelyser.

For leaf transcriptome analysis, we used three biological replicates. Each replicate was sampled from three vegetatively propagated plants per clone planted along the raw in the vineyard next to each other. Each biological replicate was separately processed during the steps of RNA extraction, library preparation, sequencing, and data analysis. Each replicate consisted of a mixture of the most distal leaves along the shoot, from the first leaf under the shoot apex to the fifth leaf. 'Pinot Meunier' was the only clone unavailable at the site of sampling, and it was collected from plant held at experimental farm of the University of Udine on the same day as the other clones.

For berry transcriptome analysis in clones of 'Sangiovese', berries were sampled at two developmental stages, before ripening (2 weeks after berry set) and at the inception of ripening (80% of coloured berries over the clusters). For each clone and sampling date, we collected three biological replicates of 30 berries each. Berries of each replicate were collected by random sampling of three berries per plant from ten plants in a row of clonally replicated vines on both sides of the canopy in north-south oriented rows. Each replicate was collected from a different plot of ten plants

in a row. Intact frozen berries were ground in liquid nitrogen. Powdered berries contained skin, flesh, and seed tissues.

VII.2 DNA-seq

DNA was extracted from nuclei and fragmented following the Illumina library preparation protocol.

For 'Sangiovese' clones, DNA was size selected by gel electrophoresis in the interval of 150-350 bp and 400-700 bp. For each clone, two libraries with fragments of different size were prepared separately and run on a Illumina Genome Analyzer II (GAII) (small-size fragments) and a Illumina Hiseq2000 platform (small-size fragments). Paired-end reads were obtained from both termini of the fragments, and the reads were 75 bp long from the GAII 100 bp long from the Hiseq2000.

For 'Pinot blanc' and 'Pinot Meunier', DNA was size selected by gel electrophoresis in the intervals of 400-600 bp and 400-700 bp. For each clone, two libraries with fragments of different size were prepared separately and run on a Hiseq2000. Paired-end reads were obtained from both termini of the fragments, and the reads were 100 bp long.

For 'Pinot gris', DNA was size selected by gel electrophoresis in the intervals of 400-700 bp and 500-1000 bp. For each clone, two libraries with fragments of different size were prepared separately and run on a Hiseq2000. Paired-end reads were obtained from both termini of the fragments, and the reads were 100 bp long.

For 'Pinot Noir', DNA was size selected by gel electrophoresis in the interval of 500-1000. A single library was prepared and run on a Hiseq2000. Paired-end reads were obtained from both termini of the fragments, and the reads were 100 bp long.

The raw reads were processed for adapter removal, quality trimming and filtering for organelle DNA and duplicates. Post-processed paired-end reads longer than 50 bp were aligned to the reference genome of PN40024 using BWA (Li and Durbin, 2009) with default parameters. Local realignment around indels was performed with the RealignerTargetCreator and IndelRealigner tools of the GATK package,

version 2.1-13 (McKenna A et al. 2010). Variant positions were identified using the UnifiedGenotyper tool of the GATK package with default parameters.

Depth-of-coverage was analysed using DNAcopy. Window width was variable along the chromosome and set in order to each window to accommodate 1,000 simulated 100-bp mappable reads, which were generated *in silico* from random fragmentation of the reference assembled sequence. Windows containing a significantly different number of normalised reads mapped from a pair of individuals were segmented by DNAcopy and the average log2 ratio of the number of reads mapped from each individual was given to the element. Paired-end mapping was used to investigate small size deletions using BreakDancerMax. A custom pipeline developed at Institute of Applied Genomics was use for the detection of the insertions with respect to *Vitis vinifera* PN40024 reference sequence. The pipeline detects the insertions resulting from known DNA elements, such as transposable elements; it is composed by three main steps: i) Putative insertions are recognized by the presence of singletons divided into two groups with opposite orientation pointing toward the putative site of insertion; their mates are expected to be unmapped because they derive from the inserted sequence; ii) the unaligned mates of the singletons are *de novo* assembled to reconstruct the two ends ('forward' and 'reverse') of the putatively inserted sequence; iii) to characterize the whole inserted sequence, the contigs obtained were aligned using blastn against a database of known plant transposable elements with the addition of the sequence regions that were identified as deletions. Insertions were detected when the two contigs aligned at the two extremities of the same sequence within this set.

VII.2.1 SNP confirmation by capillary sequencing

The DNA samples collected from leaves were the same used for NGS sequencing. Genomic DNA was extracted from 'Sangiovese' pollen with PowerPlant ® PRO DNA Isolation Kit (MO BIO Laboratories) following the manufacturer-supplied protocols and reagents.

Primer design was implemented with Primer3Plus software (Untergasser et al., 2007) based on PN40024 genomic sequence (Jaillon *et al.*, 2007). Flanking regions are between 300 bp and 500 bp long. Each couple of primers matches uniquely against reference genome PN40024 and we guarantee the absence of SNP or INDEL in primer sequences. If SNPs are present, the primer was designed with degenerated nucleotide.

DNA amplifications were performed in 15 µl PCR reactions, using KAPA2G Fast Hot Start Ready Mix (Kapa Biosystems), run in the Geneamp 9700 PCR system (Applied Biosystems, Foster City, CA), under the following conditions: 95 °C for 1 minutes, 18 cycles of 10 seconds at 95°C, 10 seconds at 70°C (-0.5°C each cycle) and 10 seconds at 72°C, 35 cycles of 10 seconds at 95°C, 15 seconds at 61°C and 10 seconds at 72°C, followed by a final extension of 7 minute at 72°C. PCR purified products were sequenced on ABI3730xl instrument according the manufacturer standard method, trimmed and assembled with Lucy/Phred/Phrap package. Each single variant was evaluated by visual inspection of pherograms with *consed*.

VII.3 RNA-seq

Leaf total RNA was extracted with a commercial kit (Sigma). RNA purity (A260/A280 nm) and quantification were estimated using a Nanodrop 1000 spectrophotometer.

An amount of 2µg of total RNA was used for library preparation following the Illumina library preparation protocol TrueSeq v2.0. RNA was fragmented into 500 bp fragments and mRNA was purified twice using poly-T beads. One library for each biological replicate was prepared, indexed and then 4 libraries were multiplexed in in a single Illumina lane. For 'Pinot' transcriptomes, 50-bp single-end reads were obtained using a Illumina Hiseq2000 platform. For 'Sangiovese' transcriptomes, paired-end reads were obtained from both termini of the fragments, and each read was 100 bp long.

The raw reads were processed for adapter removal, quality trimming and filtering for contaminants. Post-processed reads were aligned to the grapevine

transcriptome and to the reference genome using TopHat version 2.0.5 (Trapnell et al. 2012). TopHat parameters were set to map a read ten times to the reference and to report only the alignment with the best alignment score. The estimation of transcript abundance and tests for differential expression among clones were performed using Cufflinks version 2.0.2 (Trapnell et al. 2012). Significance of the difference among clones was calculated in a multi-sample run using three biological replicates per clone/tissue/sampling date. Differentially expressed genes were assigned to functional categories using BlastX and Blast2GO (selecting the output obtained for hierarchical level 2) and to metabolic pathways using MapMan (Thim O et al. 2005).

VIII List of References

Baillie JK, Barnett MW, Upton KR, Gerhardt DJ, Richmond TA, De Sapio F, Brennan PM, Rizzu P, Smith S, Fell M, Talbot RT, Gustincich S, Freeman TC, Mattick JS, Hume DA, Heutink P, Carninci P, Jeddeloh JA, Faulkner GJ (2011) Somatic retrotransposition alters the genetic landscape of the human brain. Nature 479:534-537

Bergamini C, Caputo AR, Gasparro M, Perniola R, Cardone MF, Antonacci D (2012) Evidences for an Alternative Genealogy of 'Sangiovese'. Mol Biotechnol. Epub ahead of print 2012 Mar 11. DOI 10.1007/s12033-012-9524-9

Boss PK, Thomas MR (2002) Association of dwarfism and floral induction with a grape 'green revolution' mutation. Nature 416:847-850

Cabezas JA, Ibáñez J, Lijavetzky D, Vélez D, Bravo G, Rodríguez V, Carreño I, Jermakow AM, Carreño J, Ruiz-García L, Thomas M, Martinez-Zapater JM (2011) A 48 SNP set for grapevine cultivar identification. BMC Plant Biology 11:153.

Carrier G, Le Cunff L, Dereeper A, Legrand D, Sabot F, Bouchez O, Audeguin L, Boursiquot JM, This P (2012) Transposable elements are a major cause of somatic polymorphism in Vitis vinifera L. PLoS One 7:e32973

Cipriani G, Spadotto A, Jurman I, Di Gaspero G, Crespan M, Meneghetti S, Frare E, Vignani R, Cresti M, Morgante M, Pezzotti M, Pe E, Policriti A, Testolin R (2010) The SSR-based molecular profile of 1005 grapevine (Vitis vinifera L.) accessions uncovers new synonymy and parentages, and reveals a large admixture amongst varieties of different geographic origin. *Theor Appl Genet 121:1569-1585*

Clyde A (2007) Hutchison DNA sequencing: bench to bedside and beyond. Nucl Acids Res 35:6227–6237

Conesa, A., Götz, S., García-Gómez, J.M., Terol, J., Talón, M. and Robles, M., (2005) Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics (Oxford, England), 21(18), 3674-6.

Deschamps S, Campbell MA (2009) Utilization of next-generation sequencing platforms in plant genomics and genetic variant discovery. Mol Breeding 25:553–570

Di Gaspero G, Testolin R (2013) Grapevine genomics and phenotypic diversity of bud sports, varieties and wild relatives. In Poltronieri P, Burbulis N, Fogher C(eds) From plant genomics to plant biotechnology, *Woodhead Publishing Ltd, Cambridge, UK.*

Di Vecchi Staraz M, Bandinelli R, Borselli M, This P, Boursiquot JM, Laucou V, Lacombe T (2007) Genetic structuring and parentage analysis for evolutionary studies in grapevine: kin group and origin of the cultivar Sangiovese revealed. J Amer Soc Hort Sci 132:514–524

Fernandez L, Chaïb J, Martinez-Zapater JM, Thomas MR, Torregrosa L (2012) Misexpression of a PISTILLATA-like MADS-box gene prevents fruit development in grapevine. Plant J doi: 10.1111/tpj.12083. [Epub ahead of print 2012 Nov 26]

Fernandez L, Torregrosa L, Segura V, Bouquet A, Martinez-Zapater JM (2010) Transposon-induced gene activation as a mechanism generating cluster shape somatic variation in grapevine. Plant J 61:545–557

Filippetti I, Intrieri C, Centinari M, Bucchetti B, Pastore C (2005) Molecular characterization of officially registered Sangiovese clones and of other

Sangiovese-like biotypes in Tuscany, Corsica and Emilia-Romagna Vitis 44:167–172

Franks T, Botta R, Thomas MR (2002) Chimerism in grapevines: implications for cultivar identity, ancestry and genetic improvement. Theor Appl Genet 104:192–199

Furiya T, Suzuki S, Sueta T, Takayanagi T (2009) Molecular characterization of a bud sport of Pinot gris bearing white berries. Am J Enol Vitic 60:66–73

Hillier LW, Marth GT, Quinlan AR, Dooling D, Fewell G, Barnett D, Fox P, Glasscock JI, Hickenbotham M, Huang W, Magrini VJ, Richt RJ, Sander SN, Stewart DA, Stromberg M, Tsung EF, Wylie T, Schedl T, Wilson RK, Mardis ER (2008) Whole-genome sequencing and variant discovery in C. elegans. Nat. Methods 5:183–188

Hocquigny S, Pelsy F, Dumas V, Kindt S, Héloir MC, Merdinoglu D (2004) Diversification within grapevine cultivars goes through chimeric states. Genome 47:579–589

Imazio S, Labra M, Grassi F, Winfield M, Bardini M, Scienza A: Molecular tools for clone identification: the case of grapevine cultivar 'Traminer'. Plant Breeding 2002, 121:531-535.

Jaillon O, Aury JM, Noel B, Policriti A, Clepet C, Casagrande A, Choisne N, Aubourg S, Vitulo N, Jubin C, Vezzi A, Legeai F, Hugueney P, Dasilva C, Horner D, Mica E, Jublot D, Poulain J, Bruyère C, Billault A, Segurens B, Gouyvenoux M, Ugarte E, Cattonaro F, Anthouard V, Vico V, Del Fabbro C, Alaux M, Di Gaspero G, Dumas V, Felice N, Paillard S, Juman I, Moroldo M, Scalabrin S, Canaguier A, Le Clainche I, Malacrida G, Durand E, Pesole G, Laucou V, Chatelet P, Merdinoglu D, Delledonne M, Pezzotti M, Lecharny A, Scarpelli C, Artiguenave F, Pè ME, Valle G, Morgante M, Caboche M, Adam-Blondon AF, Weissenbach J, Quétier F, Wincker P (2007) The grapevine genome sequence

suggests ancestral hexaploidization in major angiosperm phyla. Nature 449:463–467

Kovalchuk I, Kovalchuk O and Hohn B (2000), Genome-wide variation of the somatic mutation frequency in transgenicplants. The EMBO Journal Vol. 19 No. 17 pp. 4431±4438, 2000

Lacombe T, Boursiquot JM, Laucou V, Di Vecchi-Staraz M, Péros JP, This P (2012) Large-scale parentage analysis in an extended set of grapevine cultivars (Vitis vinifera L.). Theor Appl Genet. 2012 Sep 27. [Epub ahead of print] DOI 10.1007/s00122-012-1988-2

Li, H. and Durbin, R., (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics (Oxford, England), 25(14), 1754-60.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, et al. (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next- generation DNA sequencing data. Genome research 20: 1297–1303.

Metzker ML (2010) Sequencing technologies - the next generation. Nat Rev Genet 11:31–46

Myles S, Boyko AR, Owens CL, Brown PJ, Grassi F, Aradhya MK, Prins B, Reynolds A, Chia JM, Ware D, Bustamante CD, Buckler ES (2011) Genetic structure and domestication history of the grape. Proc Natl Acad Sci U S A. 108:3530-3535

Moncada X, Pelsy F, Merdinoglu D, Hinrichsen P (2006) Genetic diversity and geographical dispersal in grapevine clones revealed by microsatellite markers. Genome 49 :1459–1472

Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L. & Wold, B. (2008) Mapping and quantifying mammalian transcriptomes by RNA-seq. Nat. Methods 5, 621–628

Pelsy F (2010) Molecular and cellular mechanisms of diversity within grapevine varieties. Heredity 104:331–340

Pelsy F, Hocquigny S, Moncada X, Barbeau G, Forget D, Hinrichsen P, Merdinoglu D (2010) An extensive study of the genetic diversity within seven French wine grape variety collections. Theor Appl Genet 120:1219–1231

Petrosino, JF, Highlander S, Luna RA, Gibbs RA, Versalovic J (2009). Metagenomic pyrosequencing and microbial identification. Clin Chem 55:856–866

Pinosio, S. Building catalogues of genetic variation in Poplar (2012). Doctoral Thesis, Università degli Studi di Udine.

Ponting CP, Oliver PL, Reik W (2009) Evolution and functions of long noncoding RNAs. Cell 136:629–641

Riaz S, Garrison KE, Dangl GS, Boursiquot JM, Meredith CP (2002) Genetic divergence and chimerism within ancient asexually propagated winegrape cultivars. J Am Soc Hortic Sci 127:508–514

Schellenbaum O, Mohler V, Wenzel G, Walter B (2008) Variation in DNA methylation patterns of grapevine somaclones (Vitis vinifera L.). BMC Plant Biology, 8:78

Shendure J, Ji H (2008) Next-generation DNA sequencing. Nat Biotechnol 26:1135–1145

Silvestroni O, Intrieri C (1995) Ampelometric assessment of clonal variability in the Sangiovese wine grape cultivar. Int Symp in clonal selection, ASEV: 137-142

Soderini G. (1590). Trattato della coltivazione delle viti e del frutto che se ne puo` cavare. Societa` Tipografica de'Classici Italiani, No. 1118, Year 1806.

Stenkamp SHG, Becker MS, Hill BHE, Blaich R, Forneck A. 2009. Clonal variation and stability assay of chimeric Pinot Meunier (Vitis vinifera L.) and descending sports. Euphytica 165, 197–209

Thimm O, Bläsing O, Gibon Y, Nagel A, Meyer S, Krüger P, Selbig J, Müller LA, Rhee SY, Stitt M (2004) MAPMAN: a user-driven tool to display genomics data sets onto diagrams of metabolic pathways and other biological processes. Plant J. 37:914-39

This P, Lacombe T, Thomas MR (2006), Historical origins and genetic diversity of wine grapes. Trends Genet.;22(9):511-9

Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pimentel H, Salzberg SL, Rinn JL, Pachter L. (2012) Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat Protoc. 2012 Mar 1;7(3):562-78

Untergasser, A., Nijveen, H., Rao, X., Bisseling, Ton, Geurts, R. and Leunissen, J.A.M., (2007) Primer3Plus, an enhanced web interface to Primer3. Nucleic acids research, 35(Web Server issue), W71-74

Vezzulli S, Leonardelli L, Malossini U, Stefanini M, Velasco R, Moser C (2012) Pinot blanc and Pinot gris arose as independent somatic mutations of Pinot noir. J Exp Bot 63:6359–6369

Vouillamoz JF, Monaco A, Costantini L, Stefanini M, Scienza A, Grando MS (2007) The parentage of ‘Sangiovese’, the most important Italian wine grape. Vitis, 46:19–22

Yakushiji H, Kobayashi S, Goto-YamamotoN, Tae Jeong S, Sueta T, Mitani N, Azuma A (2006) A skin colour mutation of grapevine, from black-skinned Pinot Noir to white-skinned Pinot Blanc, is caused by deletion of the functional VvmybA1 allele. Biosci Biotechnol Biochem 70:1506-1508

Walker AR, Lee E, Robinson SP. 2006. Two new grape cultivars, bud sports of Cabernet Sauvignon bearing pale-coloured berries, are the result of deletion of two regulatory genes of the berry colour locus. Plant Molecular Biology 62, 623–635.

Walker AR, Lee E, Bogs J, McDavid DAJ, Thomas MR, Robinson SP. 2007. White grapes arose through the mutation of two similar and adjacent regulatory genes. The Plant Journal 49, 772–785.

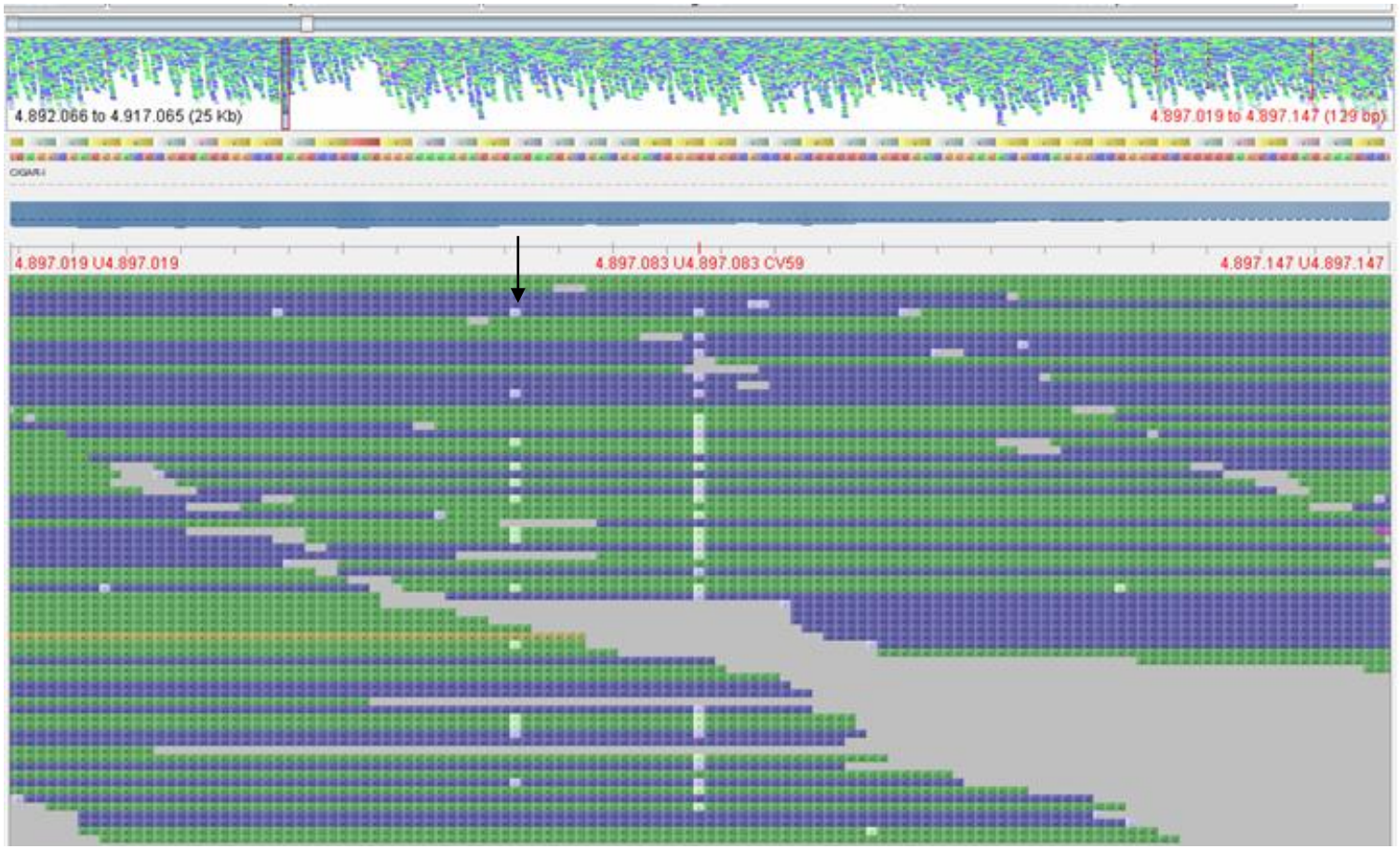
Wang Z, Gerstein M, Snyder M (2009) RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet 10:57–63

Wold B, Myers RM (2008) Sequence census methods for functional genomics. Nat Methods 5:19–21

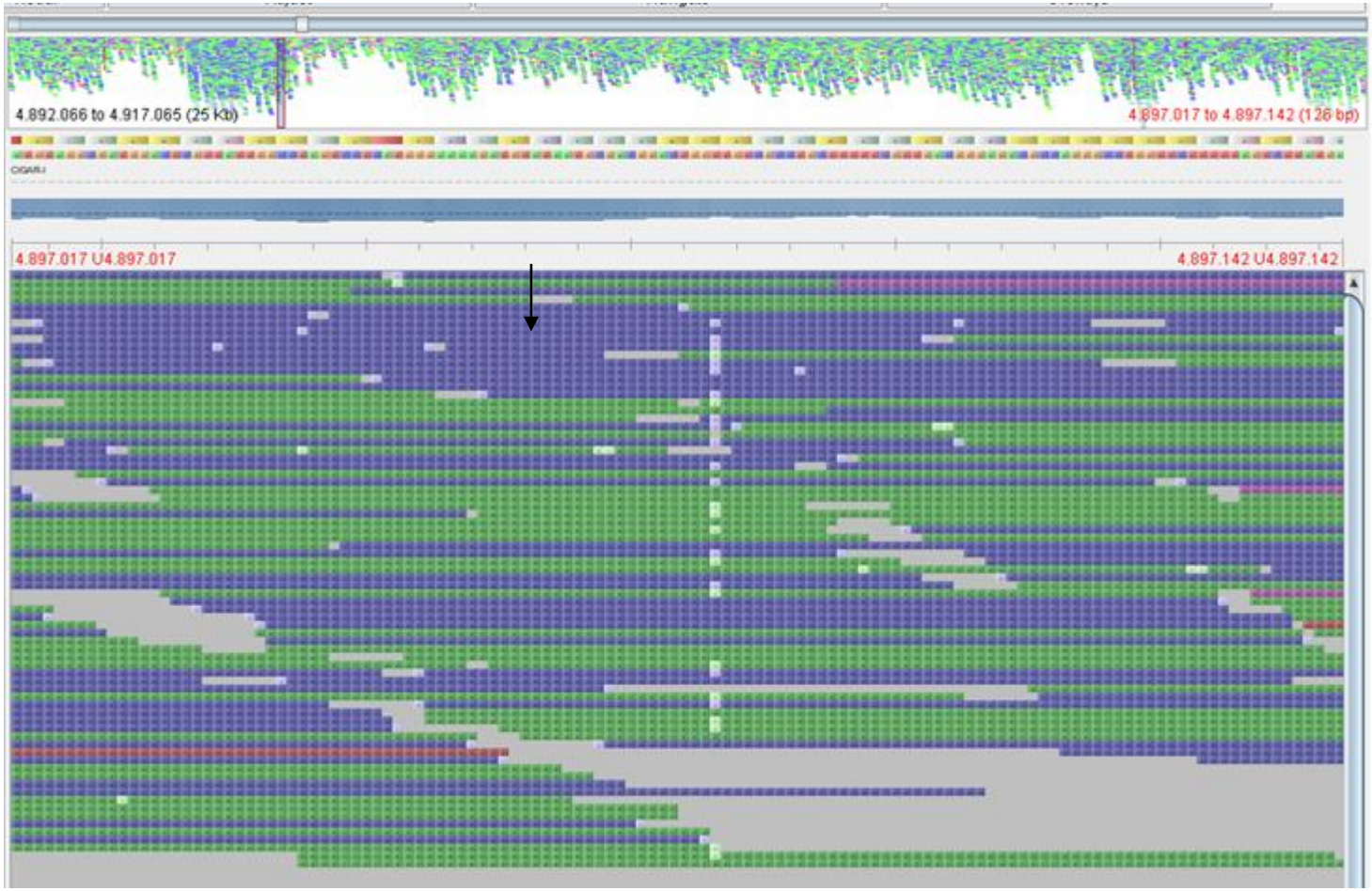
IX Appendix

Appendix 1 – Single nucleotide variant positions identified between grapevine somatic mutants

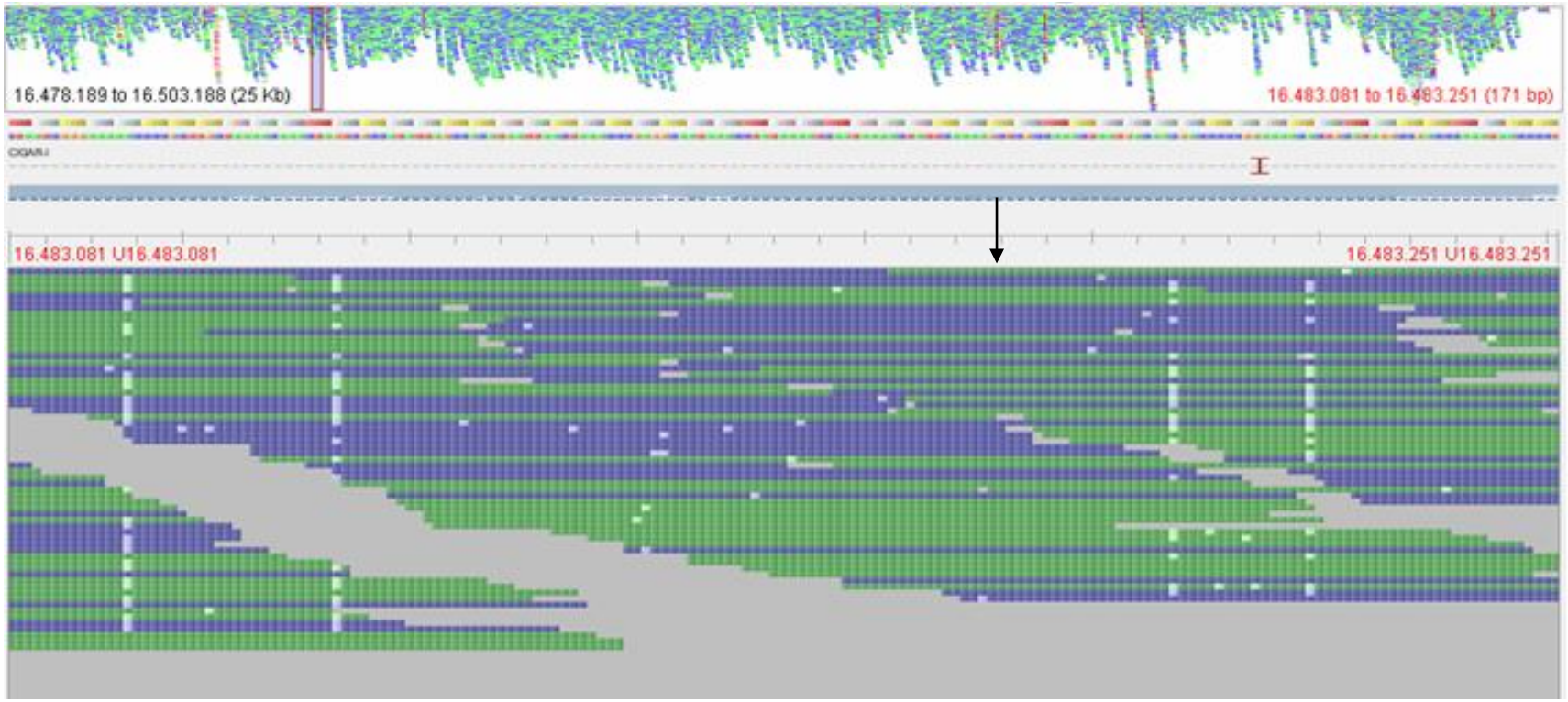
chr1:4,897,066 in 'Pinot Meunier'



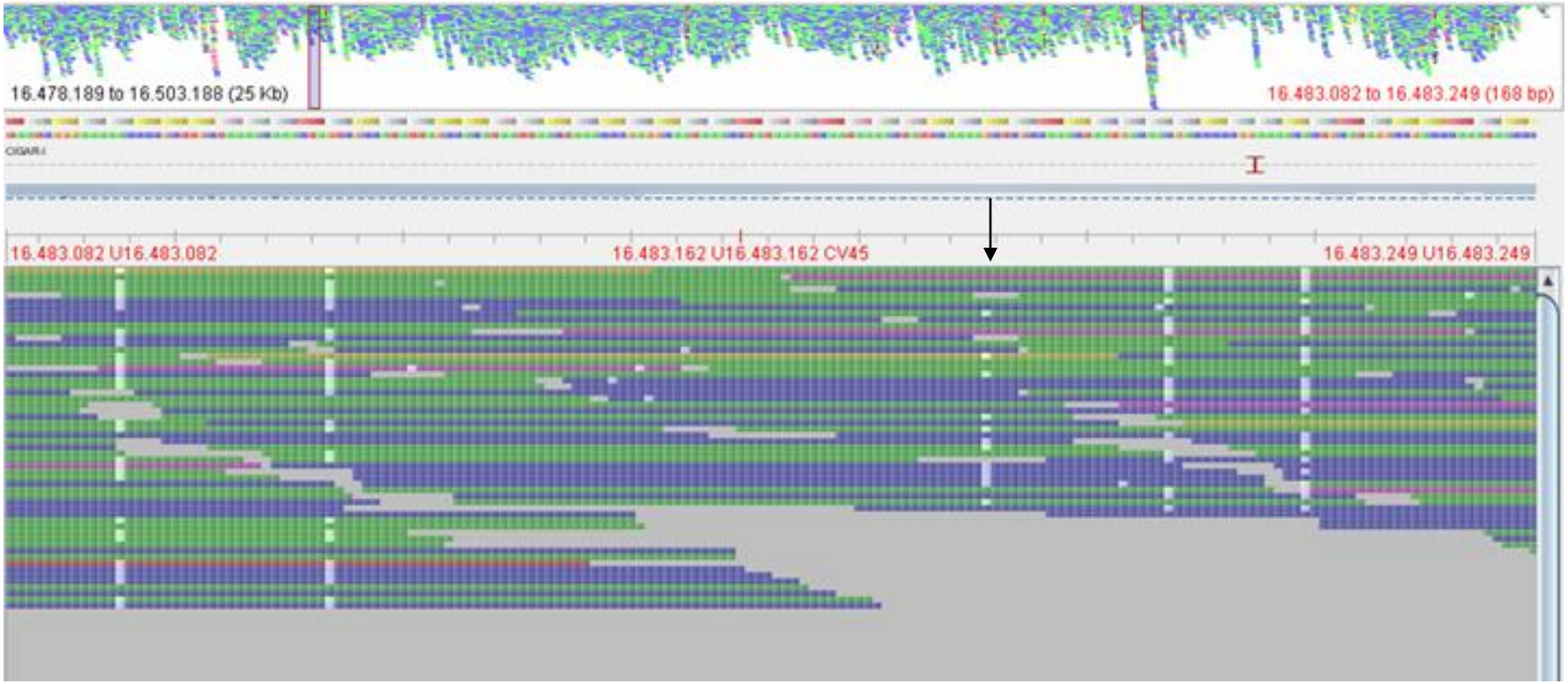
chr1:4,897,066 in ‘Pinot blanc’



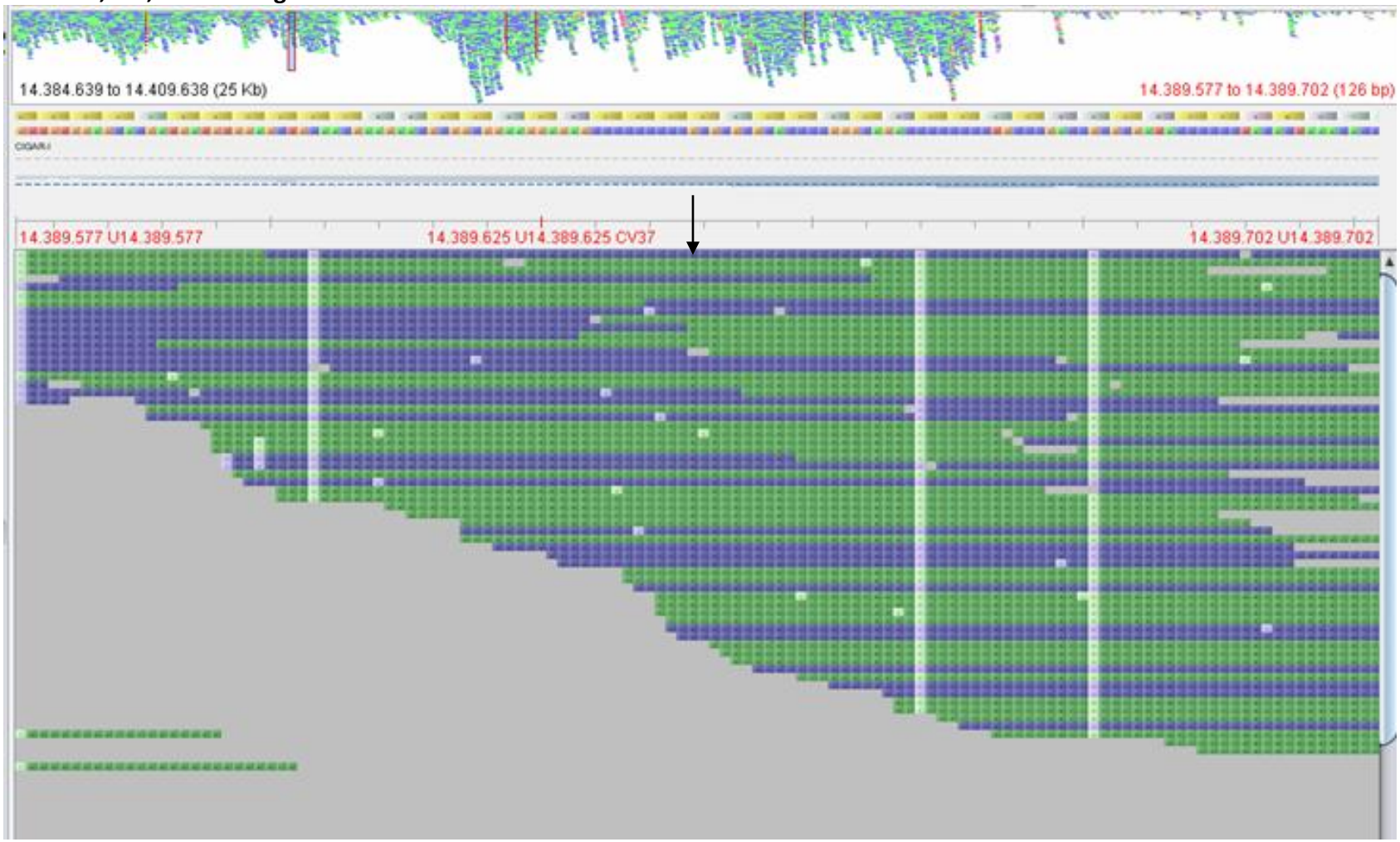
chr13:16,483,189 in 'Sangiovese R24'



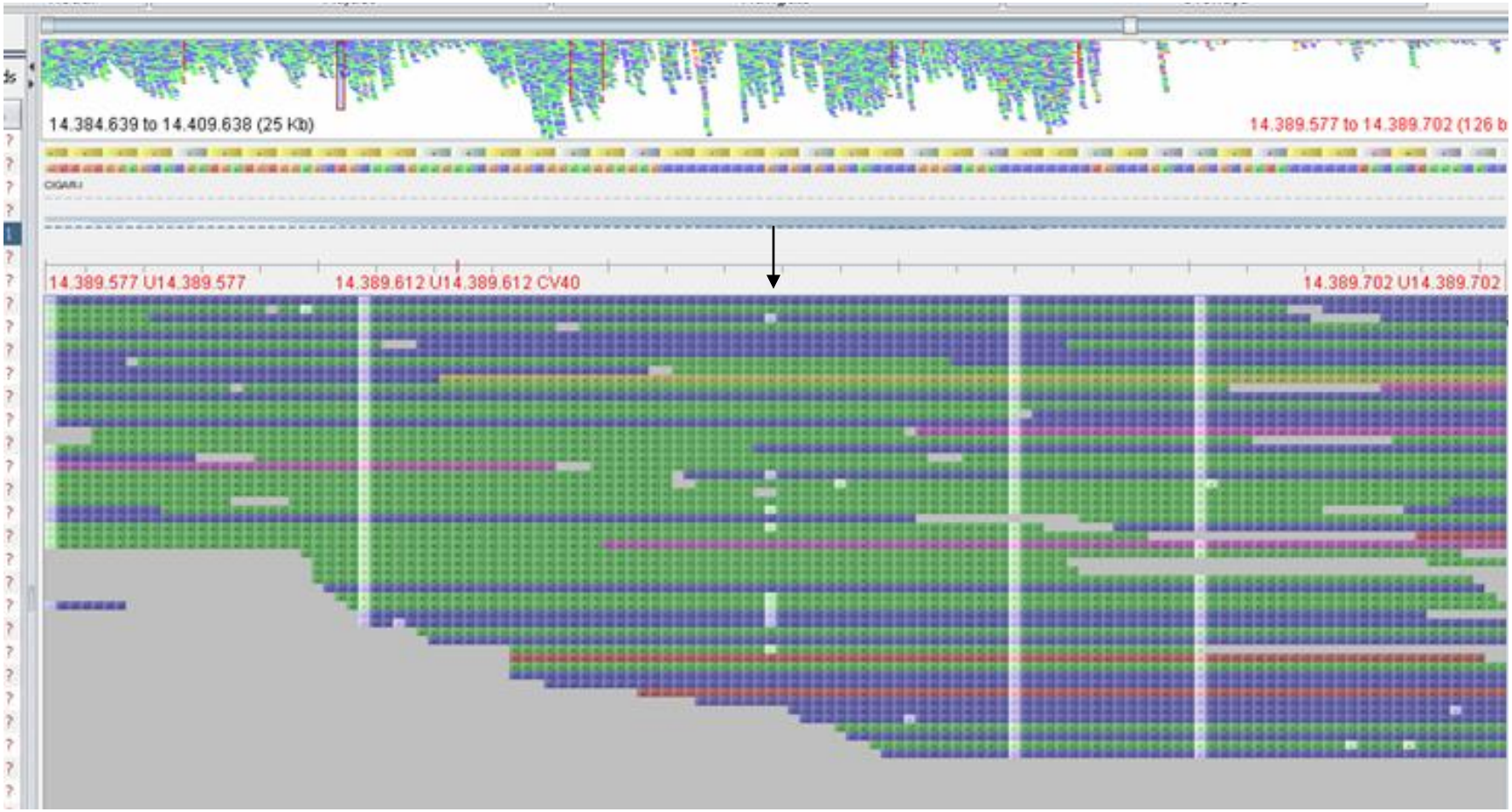
chr13:16,483,189 in 'Sangiovese VCR23'



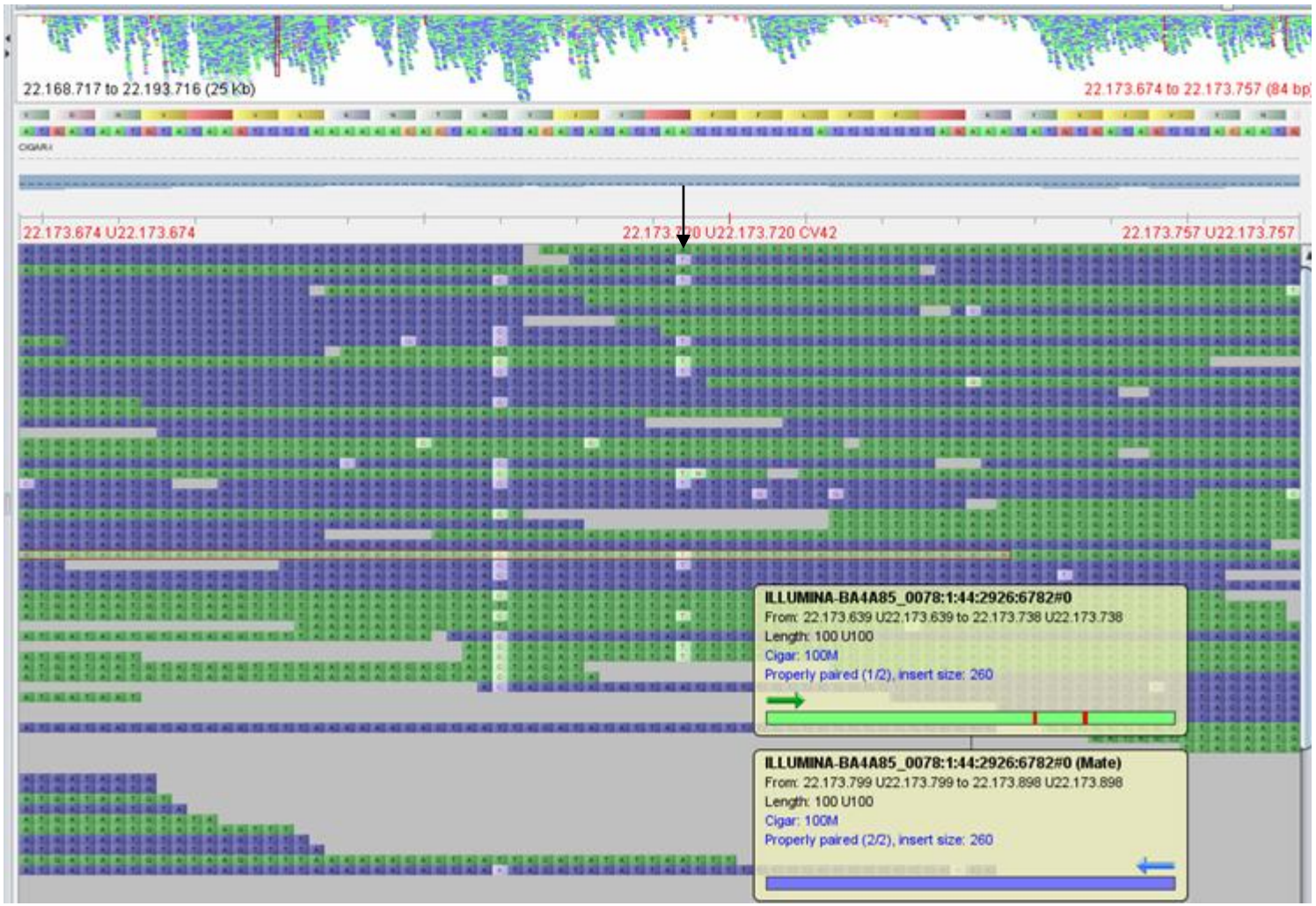
chr11:14,389,639 in 'Sangiovese R24'



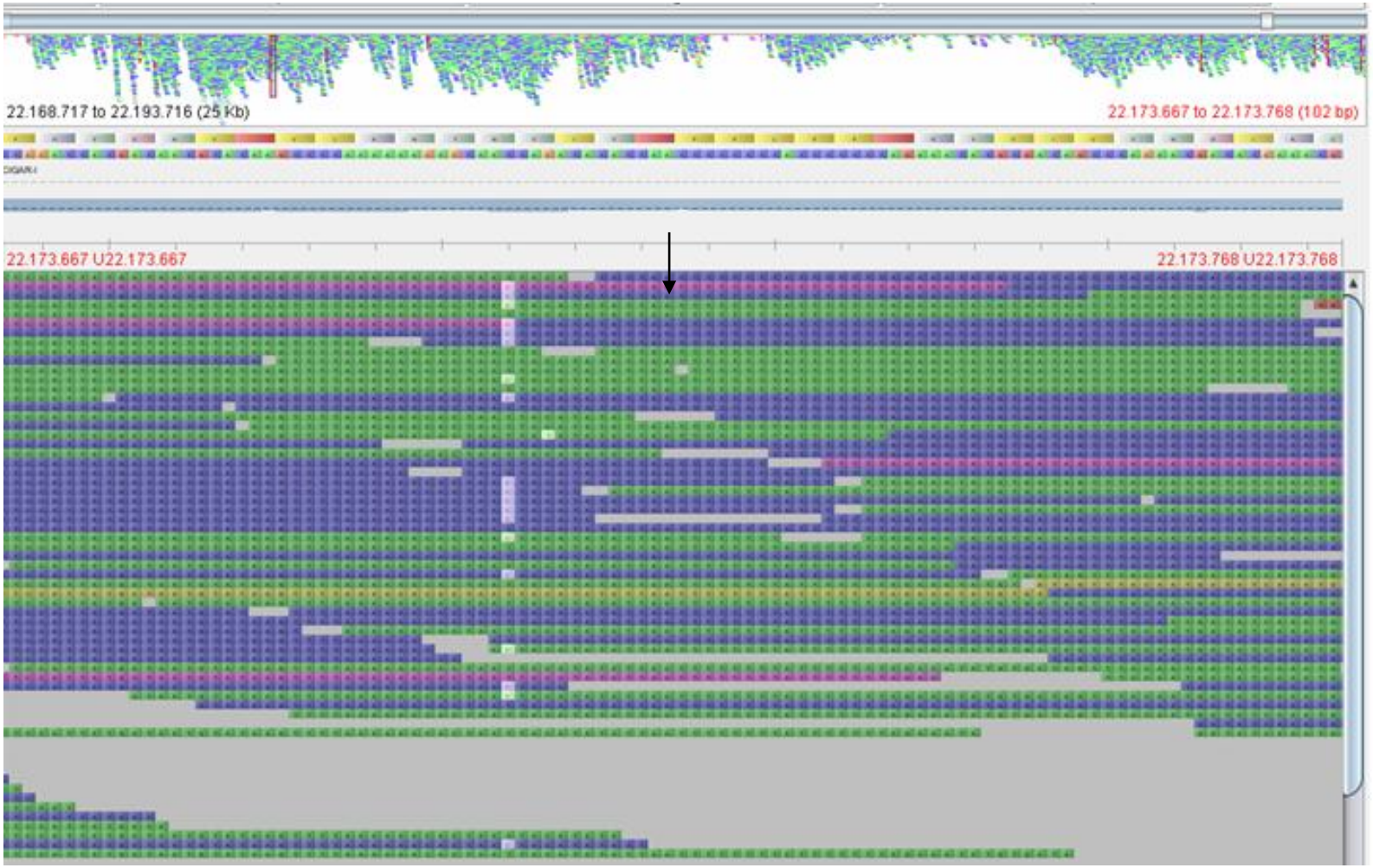
chr11:14,389,639 in 'Sangiovese VCR23'



chr19:22,173,717 in 'Sangiovese R24'



chr19:22,173,717 in 'Sangiovese VCR23'



Appendix 2 – List of primers used for PCR amplification of putative SNPs flanking region to be confirmed by Sanger sequencing

Primer_name	Sequence FORWARD	Sequence REVERSE
chr1_17035075	GCAGGAGACTGTTTGGTTGATTT	TCACTATTCTTACTCGCTTCGCA
chr1_18587330	GGCAAAAAGAACTAACGGTAGCA	A**M**CCTCACAAGCATTCAGATGA
chr1_19490545	TGGTTGGAAGGAGGGGAGAA	CAGCAAAGACAGCTCCAGGA
chr1_20214177	TCATCCCACCAAATCGTGCA	GGCTGAAAGGTCTTTGTTGCA
chr1_2995734	TTGGATAGTGGGTCATCAGAAGC	ACCATGTGAAGAAGTGGTGATGC
chr1_4745158	CAAGGTTTGGAATGGAAAGATCA	TGAAGGTTGATTTAGCTGTCGTC
chr1_4897066	TGGGTTGTCAAGGGCGCAAT	ACCCACCCTCTGCAACTCCT
chr1_5250416	CCACCCTTGAAGCCCCTATC	CCCAAATGCAGCCAACCATC
chr1_7321481	TTTCCATTAGTCTGTCCCTGTCC	TTGGAGTCGTAGCCAAATCAAAT
chr1_7458050	CACGTTGTGGTGTGGGGATA	CA**Y**GAAAGGGTCTCTTTACCAT
chr1_7545896	GGCCCTGACCCAGATTTTCA	TTGAAGCTCAGCCCTAAGCC
chr1_9398986	TTCTTGGGTTGTTGGGCTTC	AGACGTAAGCGCTCGCC**K**A
chr2_18336309	GATCCCACCAATCACCTAACCTA	AGGGCAGTACCATTTGCTTCTC
chr2_18594509	GCAAACTCACCAACAACATGAAT	CACCCACGGATAGAGAAGAGAGA
chr2_8976579	CATTGCTTGTGGGTATTTCTTCTC	CTCTACTTGCATTGTTGCACACG
chr3_642320	TTCAGCAACCCAGAGAAAGTAGC	TTCCCAGGCTATCATCCAAGTTA
chr3_7463543	ATGCAATGGTATGTTCAAGGTCA	ATCGGAAGCCTATATTCCAACAA
chr3_9302352	AGACGACAGATTGCAGATCCA	AGACGACAGATTGCAGATCCA
chr4_11645436	CACAGTATGATTTCCATTGCCCT	TACAAGACAAGTGCTGGAGGAGA
chr4_14367939	TGCTCTTCGGATGTCCCAAC	TGTTGCTATGCAGGCCTCAA
chr4_17373270	TGGTTTGTTGTCAAGGTGAAAAC	TCCCAAAAGGGCAAGCTAGT
chr4_22792206	GAAAGACTCATCATCTCCACCGTA	TTTAACCCATTTCTCTCGTGCAG
chr4_2484019	CAAATCAGAAACAACACCCATCA	TGACTTACACCCTCGTACCCAAT
chr4_2778294	TTAATCTGCAACGCCCACCT	AGGAATTGAAGCGACTGCCA
chr4_3313620	TTTCAACTTCCTTCTCCATCAGG	AGAGCCCTTACCCACCCTTATTT
chr4_3338837	CAGACTCAATGGAAGGTGGAAGT	TAAACCCTCGTCTTATGG**R**TTGC
chr4_5377597	GCTTGAACAGCAAGATGATTGAG	ACCCATCACAAACTGGCATAGAT
chr5_10019955	AATCCAGCCACAACTCAGCC	GGCTGTGAGAAGCATCTGGT
chr5_15644060	GCAACTTTGTGGTGCTGATGTAT	TCTATTGACTTCCAATGCTCTTG
chr5_15845840	TTAAACCTCCGACATGGCCC	GTTGGAGGCAAAAGCAGTGG
chr5_16947037	AAATTGTGCTCCGTGGGACT	TGCATTTCAATAGTTGAGCCACA
chr5_1911407	AGATCACCGACCGAAAGATTATG	AGCATGATTGCCTGTCCAAGTAT
chr5_19675258	AGTTGGGTTTCTCTAAACGACT	GGGTCATAGTGTCTCTTGATGTC
chr5_22745578	TACTGCCTAGGAGGA**Y**GAGC	TGAAGAGGAGCGATTGCCTC
chr5_23597312	TTGAGGTCTTCCTTTGGCATTTA	ATCTTGACTCGATAGGGCATTGA
chr5_24754569	CAACGTACCCTTAAACAACCCAC	TACTTGGCTCCTTCCATCCCTAT
chr5_4492371	GTTTCAAGATAGTTGGACCGTGG	AATTGCATTCAGCTTCACCTTTG
chr5_4735149	AGGAAGACCACTGCAGAACG	GGCTTCCCAAACAGGCTCTT
chr5_6218837	TTTCAGGTGTTTACTCCCTCACC	CAGTGCCAAGGAAAGAAGACAAC
chr5_8535969	TGGTGACACTATCCATGGTCA	TTGACCACCCACAATGGCTT

chr6_16757889	TACGCAACTCCACACCCTTC	AGTAGGTGGTGCCAATGTGG
chr6_20836866	TGATTTTGTTGCAGGCCAGC	ACGATGGAGTGCCCTCAAAG
chr6_2596937	TCGAACCTGATTATTCCTAAACCC	AGGTGGTGTTATTGGGTAGCACA
chr6_3910591	ATGATGGGAAACCTTTGGCTAAC	GGGTGTTCTAGGAAAGCACAGAA
chr6_551833	CACCACAGTTTGGCCTCTAG	CATCTGTCAGGGCCCAACTT
chr6_7669464	TCTCTACCATCTCAGCCGTCAAC	AGTGTGAACAGAAAGCGACCTCT
chr7_11468043	TGATAACACACGACATCCTTAGGT	GGAGTGACATGGGCTTTACCA
chr7_18709261	TGTGAGTTGAAGAAGAGGGCA	TGGAAACAGTCGCACAATGG
chr7_19629893	CAATCCCAGAGGAGAGATGAATG	CTGGTTGAATAAGGGAGACAACG
chr7_5802837	TGAGCAAGAGTAA**T**CACCCG	GGCACGTCTTCGCCTATGTA
chr7_6875279	TTGTGGAGCTAGGATTTGGATCT	CATCGCTACAAGAAGCCTGAAAT
chr7_9874978	ACGATTTCCCAACTACCGGA	CCTGTGTTGTGCTTAAGAGGT
chr8_11609104	AACAAATATCGCAGGGGGCA	ACACACATGGTTGGGTCAGT
chr8_14163314	AGTGCATACCTTCAACCATGTGA	TTTGGTGCTCTGTGTGCTGTAAT
chr8_15009839	TATTGTCACTGCGGAAGATACCC	TGCTCATGCTGACTCTTTGATCT
chr8_20690777	ACTCAGCAAATCAAGATACCCCA	TTGATGAGGCAGACCGGATG
chr8_2094998	TGATCCAGAATAGGCATAAGGGA	TTCTAATTGGGAAGAAAGGGCAT
chr8_3943461	TCTCTTATGGTTGGATGGCTGAT	CCATTGTTGTGGTTCTTCAATTC
chr8_454197	AGCATCAATCAGGAAACTAGGCA	GCAAGGATGTTCTATCTGTTGGC
chr8_5365582	GAAGGTCGAGGCTTGGGAAA	TGTGAGTCTCTCGGACCCAT
chr8_7696363	CGCCATTAAAGACCCAGATACAG	TTCACAAACAGGTCAAGAAACCA
chr8_8623200	GACCTTCAGTCCGTGTGTCA	TAGCTACCTCGGCTGCTCTT
chr9_18744683	GTGGGTTGAAGTCGTGAAGAAAT	GCAATCAAACGAAGACTAAACGG
chr9_3038106	CTCTCCACTGCTCCATCTGTTCT	GATGAACTTTGGTGGGTGGTTT
chr9_8342779	CGTGAGGTAAATTATCCGCTGA	ACCCAGTGTGGATAAAGCGG
chrUn_15915332	TCTCTGTGCTGGACTTCGATACA	GGACACGATAAATGTGGGTTGAG
chrUn_17599622	TCTAGCTTTCTTGGTGGATCTGC	AGGTTCATTGTGTTGGGTAACGT
chrUn_17815675	TTAAACAGCACGCAAGACAAGAG	ATAATTTGAGCAGTGGGACAAGC
chrUn_34747330	CCTAGAACATGCGCGAAGGA	AAGGTAGGACGGCTTTCCAG
chrUn_39196308	CAAATTCCCAAGCCCAGCAC	AGCCTGAACCTGTTCCTTCA
chrUn_39908858	GGAATACTGCAAGAGGGACAAGA	GGTTCAGGTAGCAATGACGTGTA
chrUn_40717178	ATAGTGGCCGCATCAATCTAGTG	TTGAATATCTCTTCCCTCGTCCT
chr10_10780961	TTTGGTTGCTTTCCTATGTGACT	ACGGGAACGAAGTTGGAGTTTAC
chr10_11534419	TGAGAAATACATGCACTCTGACCC	GCATAATAACCACTTGGCTTTGG
chr10_17545319	CCGTCCGACACTTCTCAGAC	ATCCAGTTCCTGTTGAGCCC
chr10_3531436	ATGGTTCACATGCTTCGGGG	TGAAGTGCAGAAGTGTCTCAC
chr10_4630639	TCATTTGAAAGCACTCAACAACCA	CCGTGCTTTGCCAAAGATCC
chr10_7060090	TGCATCAAGTCACCATACGTCTT	TATATGCTCCAAGGATGACCCAC
chr11_14389639	GGCCCAACTTTCCATACCCA	CC**R**AATTCACGCCCAAAACA
chr11_14389639_2	CGCCCACTCACAGCCTTCAT	ATAGAGTTGAGCGGCGTGGC
chr11_16178863	ACCCGACATGA**M**ACTGTTCC	CAACACCATTACCACACC**R**C
chr11_19240372	TTCACTACACTTGCCTCATTGCT	CATT**Y**GTTTGGTGTCTTGGAACA
chr11_4106719	ATTCCCATCACCTTTCTCGCTAT	TGCAATTGATCGATCGCACG
chr11_5376149	TAGAATGTTGCATGGTGGGTAGA	AT**Y**TGTCCTCTTGACGCGAAATA

chr11_6083181	CCTATAAATGTTTCAGCGTGGGA	GGCATAAATCCAGCAAG**W**ACAGG
chr12_11707762	AAAGAGGTCAAACAAAGTAGGTCG	ATCAAGAAACACACTTACAAATGGA
chr12_14358349	AAGAAGGTAGAAAGGTGGAACCG	CAACCTAACACAACACAAAGCCA
chr12_1638282	GAGGAAGAGAAGGTGAAGTTGGG	TGGATTGCTGTGGTATTCATTTG
chr12_17004139	AGTGATGCAGTTGTTGCTGTTGT	AAGTCCAATTCCCACTTGATTGA
chr12_2388225	GCCAGCTACCTATGGACGAC	GACGCATCATCACCTCAGCT
chr12_5656759	TCTCAAGGCTGCTACCGTTG	TGGTGAAGTGGCAATTGGGA
chr12_random_576542	AGAAGCTAACCAGAGACCCAAGG	ATGGCTCCACTCCTCCTAAAGAA
chr13_16483189	ACACCCACTAA**R**TGAAGTCCCA	CCAACTGCATGCCTTTGGTT
chr13_16483189_2	AGTCCCAAAAATTTTCTGGCTGA	CCAACTGCATGCCTTTGGTTCA
chr13_18027591	TGGAGTTGACCTGCAAGCTT	TTCCCTCTGTCGTGGTGATG
chr13_18239780	GGGCATTGTTGGCCTATATTAAAC	TACTTTCATTTGCCTTGCCTCAC
chr13_18239780	GGGCATTGTTGGCCTATATTAAAC	TA**Y**TTTCATTTGCCTTGCCTCAC
chr13_21288971	GGTTGTGAAACCGATTGGGC	GCGAGTGGAAGGGTGTCAAA
chr13_2280849	GACTGAACATGAAACCCACCTTC	GGGAGACTGATTGAGAAGAATGC
chr13_5937445	CCTAGCATATAGCGGGTAGCTCA	ACAACATTCCTTGATTGCAGACC
chr13_8600261	TCTTGAGAGGTGAGTATCCCAGC	GATGAGCCATAAGACAGTTTCGG
chr14_10718368	CAAGACATCAGGCAATACAGCAG	TCATTTGTTTCAAGATCGCTTCA
chr14_11533496	TTCGTCCCTTTAATTCTTTATTTCC	AACTCATCTCTCCATCCTCTCAAA
chr14_22875026	TGCATCCCAATGATCCAAG**R**C**R**	TGTGGCATCGAATAGTCACCA
chr14_24633241	TTCTTCAACGGTCTCATCATCCT	ATGGAGTGGACTGGACATACACC
chr14_255840	TGGGCCTGTTAGCTGTTTAGAAG	GTGGGAGTGTCTCGCTACTTGAT
chr14_2754202	AAGCCTTCCTTACGGGCTTC	TGGCTAAATGTTAGGATTATGTGGC
chr14_27794944	CTTCTCAAGGAACATTCATGGGT	TCTGATAAATCCCAAGTCCAAGTTT
chr14_27958153	GGAAGCATTTCTCACCTTTCCTT	TGACGGATTTGGTCTCTTGATTT
chr14_4034533	ATATTGGGCATCAACCTTTCCTT	CCTCCTCCTCCTCTTTCTTTCTG
chr14_4261604	TTCATTTGAGGGAACAGTGAAGAG	TAGCAGCAAAGGTGGAATTGTCT
chr14_6448337	TGTCAAATCTCTCGTCCTCCAAT	GACTCTGTCTGTCCTGTGTGCG
chr15_10456989	ACGAGTTGCAAACTATTCCACA	CATGCCAACGACCAAGACAG
chr15_11999866	TTCATACACAACCGCAAGACAAA	TGTAATGTTGAGAGAGCACAAGAA
chr15_15914659	GAAATCAATGGGATTCAAAGCAG	CAAGGGAGAGAAGTGTCGTGTTC
chr15_19571064	TGTTCTTCACCGCCTTCCAA	ATTTCAGGCCCCTTGTCACC
chr15_6384633	ACTCATTTAGGCAATCCCTCCC	AGGATGCATAGTGCCTCTGC
chr15_7340267	CAAATATGGGAAGTTGATGTTGC	TTGTATGGTGAGAGGTGAT**R**GTG
chr16_14117692	GCTTGTCTTTTGGCGCCAAT	AACCCTCAAGCAACCTCAAG
chr16_19659567	AATCAAAGCCTGAACTCCTCCAT	ATT**R**AGGTGTCAAATGGTGCTGG
chr16_21064696	GCCTGCATACTGGATTAA**K**ATCG	TCTGCATTTGGTTATTGATGTCG
chr16_2823996	GTGGACAGGAAACAACATCACAA	CATGCACTAATCTCAACCCAAGA
chr17_4282332	AGAGATGGAGGGGCACACTA	GCCTCTGGAACTTGCTCAGT
chr17_8703280	GAGAACGAAGGAGGATCTTAGGG	CCATGTTGCAGAGTATGAGCAGA
chr18_11053218	TCATTCGGTTCTGCCGTTAGTAT	ATGTCGAGGTTGCATAGGAAGTC
chr18_12490480	TAATGTCAGGCTTGTGTAAGGCA	GAAGGAGAGAGGAAGTTGCAGAA
chr18_14123317	TT**K**ACTATTTGGTTACAAGAGGGTG	TAAATCCATCA**Y**CGATCACCACA
chr18_1791435	TTAATTGATCTGTCGCGGAAAGT	GATGCAAACATCTCCAATCGAAC

chr18_19156401	GGTCTGTCTCTGTCTGGGATGTT	GGCTAATACCTTTGTCTTGCCCT
chr18_19911528	CCTCATCACAATCCGGTAAACTC	AACATTGGAAGTTGGCTTCTTGA
chr18_20694373	ACCCCACCAATCGATTGACC	CCACCACCAACAATTGATAACCA
chr18_21103474	CAGGGTATCGCAGACAGCAT	TTCCCGTCTCCAGAACTCCA
chr18_23937139	ATCAGGGTGCGACTCTAGGAAA	GTGTTGTGCCTAAACTGTCAAGC
chr18_24888322	AGTGAATCATTGTCCTCCTTGT	GTCACAGTCAGGGTGCACAG
chr18_26495835	TGAAGAACGACTTTCTCGACT	AGGGCTGAACAGTTCAATGT
chr18_28693529	GAGGCCATGAGATTGAGAATTTG	ATTGAAGAACCAACACCACCTTG
chr18_29047427	CCACAACTGCATCACTAGGAGAG	TGGGATTAGCTGCACATTTCATA
chr18_4383633	ATGCGACAGAAGTCATGCAGATA	TTGATGGGAAGGCTGAGAAGTAG
chr18_5365998	CGCATAATCATAAACCCATCACA	AAACACAAGACCTGACTGCATGA
chr18_5632015	GGTTTCTTAACCCGCTCCCA	TGTCCGAACACACTCTCTGA
chr18_5713155	ATATACCCGCCTTTCATCCACTC	TGTTGATCGAGGCAGTTTGTTT
chr18_990499	TCCTCCCAAGATCACCCTCTTAC	TATACGCATGGATGTATTGCAGG
chr19_11814802	GAAGGCGCACAAAAGTGTTG	ACTAGGGACAACACGCTTGG
chr19_1365673	GGAGAGTGGTGTGTCAAAGTTCC	TGGAGGCTTGAAAGAGAGATCAA
chr19_17977220	GAAGAACAATGAACCATCTGCCT	CAAGTAGTAATGAGCCACCAGGG
chr19_22052655	ACACTCACTCGATGCCAGAA	CAAGGGAGCATTGGGGGAAA
chr19_22173717	TCACATGCTAGACAAATTGAAAAGT	TCGAACACCGTTTACCATCTTGA
chr19_23182339	TGAGGCCAGCTAGGGAATCT	ACCCCAAAAACCAAATTGCCA
chr19_3359686	TGGTGGAGATGCTAAGTGAGT	AGTGGTGCATGTTGAGCTCA
chr19_3781226	CAAGCATTTGTTTCACTTTGGTG	ACTCACTTTCTTGCATCTGGGAA
chr19_453661	TATCAGGAACACCAGCAAGACAC	CCAAATCAAAGGGCAATACCATA
chr19_7124786	GAGTGATTTAGACCCCCTGCA	TTTGCCTGAATTTTGGGCCG
chr19_895536	ACTCATCCGGGTTCCAATAGT	TGCAATGAACCCAAACAACTC

X Acknowledgements

First of all I would like to express my sincere and deeply gratitude to my supervisor Prof. Michele Morgante for giving me the chance to work on this project at the Institute of Applied Genomics. His wide knowledge and his way of thinking have been of great value during these years.

I do would like to express my special and warm thanks to my co-supervisor Dr. Gabriele Di Gaspero. His help, support and personal guidance were really precious and fundamental for me during this work.

I would like to thank Dr. Federica Cattonaro for her patience and for the opportunities that she gave me at the Institute of Applied Genomics.

My sincere thanks to my colleague Sara Pinosio for having shared with me her capabilities on structural variation detection.

I would like to specially thank my colleague Fabio Marroni for his help and his constructive advice for this present work.

I am very grateful for the all the lab people of the Institute of Applied Genomics that taught me NGS libraries construction and helped me in the experiments, especially Vera Vendramin, Eleonora Di Centa, Nicoletta Felice, Emanuela Aleo and Irena Jurman. Their support was important throughout this work.

Additionally my thanks go to the (bio) computer scientist of Institute of Applied Genomics (IGA) Simone Scalabrin, Cristian Del Fabbro and Federico Giorgi for their kindness in answering my “bioinformatics” questions.

I would like to thank system administrator Alessandro Gervaso for his constant assistance.

A special thank to my colleagues Vera Vendramin, Vittorio Zamboni, Stefania Giacomello (IGA), Serena Foria and Simone Diego Castellarin (University of Udine). Their friendships has meant a lot for me during those years. I would like to warmly thank Simone Diego Castellarin for his patience and help during sample collection.

In addition, I have been very privileged to get to know and to collaborate with many genuinely nice people at Institute of Applied Genomics which some of them became friends over the last several years. They were able to create a great working atmosphere inside and outside the Institute, it was really a pleasure for me to work with them.

Lastly, I would like to thank my boyfriend Athos, my parents, my brother Andrea and my best friends, for their constant and lovely support. Their encouraging and their patience during the final stages of this PhD work were really appreciate. A special thank goes to my friend Marco, for having being a good listener and having supported me during all those years of work.

The financial support of Vivai Cooperativi Rauscedo (Rauscedo, Italy) is gratefully acknowledged.

Thank you.